

S/N 09/891,080
ART UNIT 2151

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 11-191135

(43)Date of publication of application : 13.07.1999

(51)Int.Cl.

G06K 9/20

(21)Application number : 10-125103

(71)Applicant : RICOH CO LTD

(22)Date of filing : 07.05.1998

(72)Inventor : MIZUNA TOORU
SAITO TAKASHI

(30)Priority

Priority number : 09245523
09287204Priority date : 10.09.1997
20.10.1997

Priority country : JP

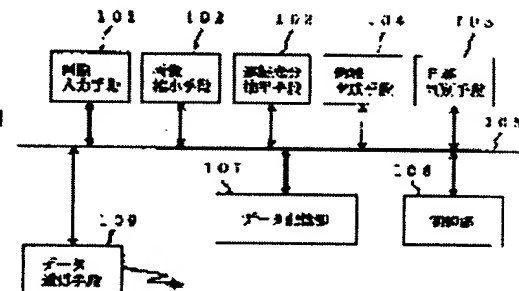
JP

(54) JAPANESE/ENGLISH DISCRIMINATING METHOD FOR DOCUMENT IMAGE, DOCUMENT RECOGNIZING METHOD AND RECORDING MEDIUM

(57)Abstract:

PROBLEM TO BE SOLVED: To accurately identify Japanese and English at high speed and to identify both the language even concerning a range to be identified for each character area and each page unit.

SOLUTION: After an input document image 101 is reduced 102, a black pixel connection component is extracted 103, and the character area is generated 104 by merging these components. Concerning the generated character area, based on the length of the connection component, a Japanese/English discriminating means 104 classifies that component and based on the accumulated value of the classified result, whether it is a Japanese area or an English area is discriminated.



Best Available Copy

LEGAL STATUS

[Date of request for examination]

11.09.2002

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2003 Japan Patent Office

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平11-191135

(43) 公開日 平成11年(1999) 7月13日

(51) Int.Cl.⁶

G 0 6 K 9/20

識別記号

3 2 0

F I

G 0 6 K 9/20

3 2 0 P

審査請求 未請求 請求項の数15 O L (全 27 頁)

(21) 出願番号 特願平10-125103

(22) 出願日 平成10年(1998) 5月7日

(31) 優先権主張番号 特願平9-245523

(32) 優先日 平9(1997) 9月10日

(33) 優先権主張国 日本 (J P)

(31) 優先権主張番号 特願平9-287204

(32) 優先日 平9(1997) 10月20日

(33) 優先権主張国 日本 (J P)

(71) 出願人 000006747

株式会社リコー

東京都大田区中馬込1丁目3番6号

(72) 発明者 水納 亨

東京都大田区中馬込1丁目3番6号 株式
会社リコー内

(72) 発明者 齋藤 高志

東京都大田区中馬込1丁目3番6号 株式
会社リコー内

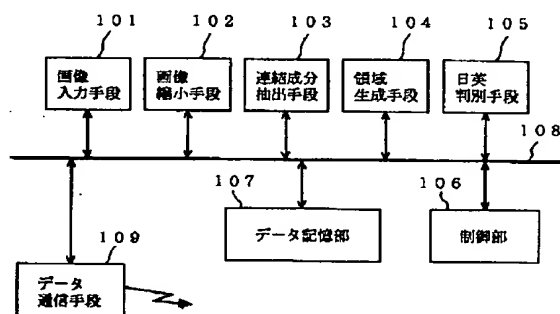
(74) 代理人 弁理士 鈴木 誠 (外1名)

(54) 【発明の名称】 文書画像の日本語英語判定方法、文書認識方法および記録媒体

(57) 【要約】

【課題】 精度よくかつ高速に日本語と英語の識別を行うと共に、識別する範囲についても各文字領域毎に、またページ単位毎に両者を識別できる。

【解決手段】 入力文書画像101を縮小102した後、黒画素連結成分を抽出103し、それらを統合して文字領域を生成104する。生成した文字領域について、日英判別手段105は、連結成分の長さを基にその成分を分類し、分類結果の集計値を基に日本語領域であるか英語領域であるかを判別する。



【特許請求の範囲】

【請求項1】 文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、複数の判定方法を用いて日本語領域であるか英語領域であるかを判定し、該複数の判定結果を比較することによって最終判定結果を得ることを特徴とする文書画像の日本語英語判定方法。

【請求項2】 文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、前記文書画像を縮小することにより生成される文字領域内の黒画素連結成分の長さを基に該連結成分を分類し、該分類結果の集計値を基に前記各文字領域が日本語領域であるか英語領域であるかを判定することを特徴とする文書画像の日本語英語判定方法。

【請求項3】 前記生成される文字領域内の黒画素連結成分の数が所定の条件を満たさないとき、異なる判定方法を用いることを特徴とする請求項2記載の文書画像の日本語英語判定方法。

【請求項4】 各ページの文書画像が日本語文書画像であるか英語文書画像であるかを判定する文書画像の日本語英語判定方法であって、前記文書画像を縮小することにより生成されるページ内の黒画素連結成分の長さを基に該連結成分を分類し、該分類結果の集計値を基に前記各ページが日本語領域であるか英語領域であるかを判定することを特徴とする文書画像の日本語英語判定方法。

【請求項5】 ページが複数の文字領域からなり、各ページの文書画像が日本語文書画像であるか英語文書画像であるかを判定する文書画像の日本語英語判定方法であって、前記文書画像を縮小することにより生成される文字領域内の黒画素連結成分の長さを基に該連結成分を分類し、該分類結果の集計値を基に前記各文字領域が日本語領域であるか英語領域であるかを判定し、該判定結果を基に前記各ページが日本語領域であるか英語領域であるかを判定することを特徴とする文書画像の日本語英語判定方法。

【請求項6】 文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、前記文字領域中から行を検出し、該行中から近接した外接矩形を統合してブロックを抽出し、該ブロック毎に日本語領域であるか英語領域であるか、あるいは判定不能領域であるかを判定し、該判定結果を前記ブロック毎に集計し、該集計値を基に前記各文字領域が日本語領域であるか英語領域であるかを判定することを特徴とする文書画像の日本語英語判定方法。

【請求項7】 前記抽出されるブロックの数が所定の条件を満たさないとき、異なる判定方法を用いることを特徴とする請求項6記載の文書画像の日本語英語判定方法。

【請求項8】 ページが複数の文字領域からなり、各ページの文書画像が日本語文書画像であるか英語文書画像

であるかを判定する文書画像の日本語英語判定方法であって、前記文字領域中から行を検出し、該行中から近接した外接矩形を統合してブロックを抽出し、該ブロック毎に日本語領域であるか英語領域であるか、あるいは判定不能領域であるかを判定し、該判定結果をページ単位で集計し、該集計値を基に前記各ページが日本語文書画像であるか英語文書画像であるかを判定することを特徴とする文書画像の日本語英語判定方法。

【請求項9】 ページが複数の文字領域からなり、各ページの文書画像が日本語文書画像であるか英語文書画像であるかを判定する文書画像の日本語英語判定方法であって、前記文字領域中から行を検出し、該行中から近接した外接矩形を統合してブロックを抽出し、該ブロック毎に日本語領域であるか英語領域であるか、あるいは判定不能領域であるかを判定し、該判定結果を文字領域毎に集計し、該集計値を基に文字領域毎に日本語領域であるか英語領域であるかを判定し、該判定結果をページ単位で集計し、該集計値を基に前記各ページが日本語文書画像であるか英語文書画像であるかを判定することを特徴とする文書画像の日本語英語判定方法。

【請求項10】 文書画像が日本語文書画像であるか英語文書画像であるかを判定し、該判定結果に応じた文書認識処理を行うことを特徴とする文書認識方法。

【請求項11】 文書画像を複数の文字領域に分割し、該分割された文字領域毎に日本語文書領域であるか英語文書領域であるかを判定し、該判定結果に応じた文書認識処理を行うことを特徴とする文書認識方法。

【請求項12】 文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定するために、複数の判定方法を用いて日本語領域であるか英語領域であるかを判定する機能と、該複数の判定結果を比較することによって最終判定結果を得る機能をコンピュータに実現させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項13】 文書画像中の各文字領域または各ページの文書画像が日本語領域であるか英語領域であるかを判定するために、前記文書画像を縮小することにより生成される文字領域内またはページ内の黒画素連結成分の長さを基に該連結成分を分類する機能と、該分類結果の集計値を基に前記各文字領域または各ページが日本語領域であるか英語領域であるかを判定する機能をコンピュータに実現させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項14】 文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定するために、または、ページが複数の文字領域からなり、各ページの文書画像が日本語文書画像であるか英語文書画像であるかを判定するために、前記文字領域中から行を検出する機能と、該行中から近接した外接矩形を統合してブロックを抽出する機能と、該ブロック毎に日本語領域であるか英語領

域であるか、あるいは判定不能領域であるかを判定する機能と、該判定結果を前記ブロック毎またはページ単位に集計する機能と、該集計値を基に、前記各文字領域が日本語領域であるか英語領域であるかを判定する機能または各ページが日本語文書画像であるか英語文書画像であるかを判定する機能をコンピュータに実現させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項15】 文書画像が日本語文書画像であるか英語文書画像であるかを判定する機能または文書画像を複数の文字領域に分割し、該分割された文字領域毎に日本語文書領域であるか英語文書領域であるかを判定する機能と、該判定結果に応じた文書認識処理を行う機能をコンピュータに実現させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、文書画像中の各文字領域に対して日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法および記録媒体に関し、また文書画像が日本語文書画像であるか英語文書画像であるかを判定してから認識処理する文書認識方法および記録媒体に関する。

【0002】

【従来の技術】文書画像に対して文字認識処理を施す場合に、適切な言語を選択する必要がある。すなわち、英文OCRで日本語を認識しようとしてもアルファベットや数字以外は認識不可能であるし、また逆に日本語OCRで英文を認識しようすると、文字切り出しや言語処理のうで英文OCRを使用した場合よりも認識率が低くなってしまふ。

【0003】従って、文字認識処理を施す前に、言語識別を行う必要が生じる。従来から文書中の文字種を識別する種々の手法が提案されている。例えば、2値化された文字行の縦方向または横方向の黒白反転回数を計数し、その分布を基に文字種の識別を行う文書認識装置がある（特開平5-108876号公報を参照）。

【0004】また、読み取った単語を認識させ、その認識結果と辞書との適合率を基に認識文字の言語種類を判別する文書認識装置もある（特開平6-150061号公報を参照）。

【0005】

【発明が解決しようとする課題】上記した前者の装置では、文字種を識別する特徴として黒白反転回数をを用いているが、この特徴はフォントや文書内容（かな、漢字、数字などの比率）による変動が大きく、このために識別の精度が低くなるという問題がある。

【0006】これに対して、後者の装置では、一度、文字認識を行っているの、OCRの性能がよければかなりの確率で字種が判明することになり、精度よく日英判

別を行うことが可能となる。しかし、OCRは処理に多くの時間を要するという問題がある。

【0007】本発明は上記した事情を考慮してなされたもので、本発明の目的は、精度よくかつ高速に日本語と英語の識別を行うと共に、識別する範囲についても各文字領域毎に、またページ単位毎に両者を識別できる文書画像の日本語英語判別方法および記録媒体、さらには、文書画像を判定し、最適な文書認識処理を行う文書認識方法および記録媒体を提供することにある。

【0008】

【課題を解決するための手段】前記目的を達成するために、請求項1記載の発明では、文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、複数の判定方法を用いて日本語領域であるか英語領域であるかを判定し、該複数の判定結果を比較することによって最終判定結果を得ることを特徴としている。

【0009】請求項2記載の発明では、文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、前記文書画像を縮小することにより生成される文字領域内の黒画素連結成分の長さを基に該連結成分を分類し、該分類結果の集計値を基に前記各文字領域が日本語領域であるか英語領域であるかを判定することを特徴としている。

【0010】請求項3記載の発明では、前記生成される文字領域内の黒画素連結成分の数が所定の条件を満たさないとき、異なる判定方法を用いることを特徴としている。

【0011】請求項4記載の発明では、各ページの文書画像が日本語文書画像であるか英語文書画像であるかを判定する文書画像の日本語英語判定方法であって、前記文書画像を縮小することにより生成されるページ内の黒画素連結成分の長さを基に該連結成分を分類し、該分類結果の集計値を基に前記各ページが日本語領域であるか英語領域であるかを判定することを特徴としている。

【0012】請求項5記載の発明では、ページが複数の文字領域からなり、各ページの文書画像が日本語文書画像であるか英語文書画像であるかを判定する文書画像の日本語英語判定方法であって、前記文書画像を縮小することにより生成される文字領域内の黒画素連結成分の長さを基に該連結成分を分類し、該分類結果の集計値を基に前記各文字領域が日本語領域であるか英語領域であるかを判定し、該判定結果を基に前記各ページが日本語領域であるか英語領域であるかを判定することを特徴としている。

【0013】請求項6記載の発明では、文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、前記文字領域中から行を検出し、該行中から近接した外接矩形を統合してブロックを抽出し、該ブロック毎に日本語領域

10

20

30

40

50

であるか英語領域であるか、あるいは判定不能領域であるかを判定し、該判定結果を前記ブロック毎に集計し、該集計値を基に前記各文字領域が日本語領域であるか英語領域であるかを判定することを特徴としている。

【0014】請求項7記載の発明では、前記抽出されるブロックの数が所定の条件を満たさないとき、異なる判定方法を用いることを特徴としている。

【0015】請求項8記載の発明では、ページが複数の文字領域からなり、各ページの文書画像が日本語文書画像であるか英語文書画像であるかを判定する文書画像の日本語英語判定方法であって、前記文字領域中から行を検出し、該行中から近接した外接矩形を統合してブロックを抽出し、該ブロック毎に日本語領域であるか英語領域であるか、あるいは判定不能領域であるかを判定し、該判定結果をページ単位で集計し、該集計値を基に前記各ページが日本語文書画像であるか英語文書画像であるかを判定することを特徴としている。

【0016】請求項9記載の発明では、ページが複数の文字領域からなり、各ページの文書画像が日本語文書画像であるか英語文書画像であるかを判定する文書画像の日本語英語判定方法であって、前記文字領域中から行を検出し、該行中から近接した外接矩形を統合してブロックを抽出し、該ブロック毎に日本語領域であるか英語領域であるか、あるいは判定不能領域であるかを判定し、該判定結果を文字領域毎に集計し、該集計値を基に文字領域毎に日本語領域であるか英語領域であるかを判定し、該判定結果をページ単位で集計し、該集計値を基に前記各ページが日本語文書画像であるか英語文書画像であるかを判定することを特徴としている。

【0017】請求項10記載の発明では、文書画像が日本語文書画像であるか英語文書画像であるかを判定し、該判定結果に応じた文書認識処理を行うことを特徴としている。

【0018】請求項11記載の発明では、文書画像を複数の文字領域に分割し、該分割された文字領域毎に日本語文書領域であるか英語文書領域であるかを判定し、該判定結果に応じた文書認識処理を行うことを特徴としている。

【0019】請求項12記載の発明では、文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定するために、複数の判定方法を用いて日本語領域であるか英語領域であるかを判定する機能と、該複数の判定結果を比較することによって最終判定結果を得る機能をコンピュータに実現させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体であることを特徴としている。

【0020】請求項13記載の発明では、文書画像中の各文字領域または各ページの文書画像が日本語領域であるか英語領域であるかを判定するために、前記文書画像を縮小することにより生成される文字領域内またはペー

ジ内の黒画素連結成分の長さを基に該連結成分を分類する機能と、該分類結果の集計値を基に前記各文字領域または各ページが日本語領域であるか英語領域であるかを判定する機能をコンピュータに実現させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体であることを特徴としている。

【0021】請求項14記載の発明では、文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定するために、または、ページが複数の文字領域からなり、各ページの文書画像が日本語文書画像であるか英語文書画像であるかを判定するために、前記文字領域中から行を検出する機能と、該行中から近接した外接矩形を統合してブロックを抽出する機能と、該ブロック毎に日本語領域であるか英語領域であるか、あるいは判定不能領域であるかを判定する機能と、該判定結果を前記ブロック毎またはページ単位に集計する機能と、該集計値を基に、前記各文字領域が日本語領域であるか英語領域であるかを判定する機能または各ページが日本語文書画像であるか英語文書画像であるかを判定する機能をコンピュータに実現させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体であることを特徴としている。

【0022】請求項15記載の発明では、文書画像が日本語文書画像であるか英語文書画像であるかを判定する機能または文書画像を複数の文字領域に分割し、該分割された文字領域毎に日本語文書領域であるか英語文書領域であるかを判定する機能と、該判定結果に応じた文書認識処理を行う機能をコンピュータに実現させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体であることを特徴としている。

【0023】

【発明の実施の形態】以下、本発明の一実施例を図面を用いて具体的に説明する。

〈実施例1〉図1は、本発明の実施例1の構成を示す。図において、101は、文書画像を入力する画像入力手段、102は、入力文書画像を縮小する画像縮小手段、103は、文書画像から連結成分を抽出する連結成分抽出手段、104は、抽出した連結成分を分類し、統合することによって文字領域を生成する領域生成手段、105は、文字領域単位またはページ単位で日本語と英語を判別する日英判別手段、106は、全体を制御する制御部、107は、入力された文書画像データや連結成分データ、領域データなど各種データを記憶するデータ記憶部、108は、データ通信路、109は、ネットワーク、回線などを介してホストなどに接続するデータ通信手段である。

【0024】図2は、本発明の実施例1の全体の処理フローチャートを示す。以下、図2を参照しながら、本発明の処理動作を説明する。まず、画像入力手段101は、文書を読み取ることによって文書画像を得る（ステ

ップ201)。この画像入力手段は、例えばスキャナ、ファックスなどであり、またデータ通信手段109を介してネットワーク経由で別の機器から画像を得るようにしてもよい。

【0025】次に、画像縮小手段102は、入力された文書画像を縮小する(ステップ202)。この処理は、例えば入力文書画像を1/8程度にOR縮小する処理である。すなわち、8×8画素を1画素に縮小するもので、64画素中に1つでも黒画素があれば縮小画素は黒画素とする処理である。

【0026】連結成分抽出手段103は、縮小画像から黒画素連結成分を抽出する(ステップ203)。領域生成手段104は、抽出した連結成分を分類し、統合して文字領域を生成する(ステップ204)。この領域生成方法として、例えば特開平6-20092号公報に記載された公知の方法を用いればよい。このとき、各文字領域を構成する連結成分の情報はデータ記憶部107に格納、保持する。

【0027】続いて、生成した文字領域について、日英判別手段105は日本語か英語かの判定を行う(ステップ205)。

【0028】ステップ202において画像をOR縮小することにより、近傍の黒画素どうしが融合する。ここで英文においては単語間にはスペースが存在し、単語内の文字間は非常に狭いという特徴がある。一方、日本語においては、句読点の前後以外では文字間隔は大きくは変わらない。

【0029】図3は、英文、日本語文の画像例と、その外接矩形を示す。英文画像301を縮小し、連結成分を抽出した結果を外接矩形で表現したものが外接矩形302である(なお、縮小処理しているので外接矩形302は、本来画像301より小さくなるべきだが、ここでは同じサイズで表現している)。英文画像では、単語毎に融合して連結成分が構成される。

【0030】日本語画像303と305の例について、同様に縮小して連結成分を抽出し、その外接矩形で表現すると、それぞれ外接矩形304、306のようになる。

【0031】英文の場合は、単語を構成する文字の数がある程度一定であるので、縦横比が2倍から6、7倍程度となる外接矩形が多くなる特徴がある。一方、日本語の場合は、外接矩形304に示すように英文では現れにくい長い矩形が生じたり、逆に外接矩形306のように細かい矩形が多く生じる特徴がある。

【0032】そこで、上記した連結成分矩形を「短」、「中」、「長」の3種類に分類し、これを各文字領域について集計する。図4は、実施例1の日英判定の処理フローチャートを示す。図4の処理は各文字領域毎に行われる。矩形の分類は、行方向が横の場合には例えば、幅/高さが2以下で「短」、幅/高さが2から6で

「中」、それ以上で「長」とする(ステップ401)。そして、文字領域中におけるこの分類結果を集計し(ステップ402)、文字領域毎に日本語か英語かを判定する(ステップ403)。ここで、「短」矩形の数をSCNT、「中」矩形の数をNCNT、「長」矩形の数をLCNTとすると、日英の判定は図8(ステップ403の詳細フローチャート)に示すように行われる。

【0033】まず、 $LCNT / (NCNT + SCNT) > Th1$ が成り立つかどうか調べる(ステップ801)。 $Th1$ は予め定めたいきい値であり、例えば0.3程度とする。この条件式が成り立てば、長矩形が十分に多いということであり、当該文字領域は日本語領域であると判定する(ステップ804)。

【0034】次に、ステップ801でNoと判定されたとき、 $NCNT / (LCNT + SCNT) < Th2$ が成り立つかどうかを調べる(ステップ802)。 $Th2$ も予め定めたいきい値であり、例えば3とする。この条件式が成り立てば、中矩形が少ないということであり、当該文字領域は日本語領域であると判定する(ステップ804)。いずれの条件も満たさない場合は、英語領域と判定される(ステップ803)。

【0035】〈実施例2〉上記した実施例1では、文字領域単位で日英の判定を行っている。この場合、文字領域によっては文字数が非常に少ない場合がある。そのような場合は、矩形の数が十分に得られないので矩形数の比率で日英判定を行うことが難しくなる可能性がある。実施例2は、矩形の数が十分でない場合を考慮した実施例である。

【0036】図5は、実施例2の処理フローチャートを示す。日英判別手段105は、集計された領域内の矩形の数が十分であるか否か(つまり所定の閾値 Th 以上あるか否か)を調べ(ステップ501)、十分でない場合には、前掲した特開平6-150061号公報に記載されているOCRを利用した日英判別を行う(ステップ503)。この場合は、文字の数が少ないのでOCR処理を施しても処理時間の増大は少なくてすむ。そして、矩形の数が十分である場合には実施例1で説明した矩形長による日英の識別を行う(ステップ502)。

【0037】〈実施例3〉次に、ページ単位で日英識別を行う実施例3について説明する。図6、7は、実施例3に係るステップ205の詳細フローチャートを示す。図6に示す方法は、「短」、「中」、「長」矩形の数の集計を文字領域毎でなくページ全体について行い(ステップ601、602)、その結果を使用してページ単位に日英の判定を行う(ステップ603)。この日英の判定方法は、図8の処理フローチャートに従って行う。このときのしきい値 $Th1$ 、 $Th2$ は文字領域単位の処理の場合と異なるしきい値としてもよい。

【0038】図7に示す方法は、各文字領域毎に日英の判別を行い(ステップ702)、その結果を基に当該ペ

ージの日英判定を行う(ステップ703)。具体的には、日本語領域と判定された領域の数を J_n 、英語領域と判定された領域の数を E_n として、 $J_n > E_n$ なら日本語ページ、 $E_n > J_n$ なら英語ページと判定する。 $J_n = E_n$ の場合はリジェクトし、あるいは日英の何れかに判定してもよい。

【0039】〈実施例4〉上記した実施例とは異なる特徴を利用した日英識別方法について説明する。図9は、実施例4の構成を示す。実施例1と異なる点は、行切り出し部902と、ブロック抽出部903と、ブロック内文字種判別部904を設けている点である。他の構成要素は実施例1のものと同様である。図10は、実施例4の処理フローチャートを示す。

【0040】まず、行切り出し部902は、文書画像の文字領域から行の切り出しを行う(ステップ1001、1002)。領域生成処理として、特開平6-20092号公報記載の技術を使用した場合には、領域を抽出した段階で行情報が得られているので、これを用いればよく、また電子通信学会論文「周辺密度分布、線密度、外接矩形特徴を利用した文書画像の領域分割」(秋山他、1986年8月、Vol. J69-D No. 8)に記載されている射影を用いる方法を用いてもよい。

【0041】次に、ブロック抽出部903は、単語相当のブロックを抽出する(ステップ1003)。このブロック抽出方法として、本出願人が先に特願平8-34781号で提案した方法を用いればよい。すなわち、ブロック抽出部111は、行データ内部の外接矩形を検出し、その外接矩形をブロックデータにまとめる。このブロックデータにまとめる方法は、次の通りである。文字矩形の間隔(まだ一つの矩形が一字とは確定されていない。従って、漢字の場合、偏とつくりに分離したものがそれぞれ一つの矩形となる場合も多い)のヒストグラムを求める。図18は、抽出された文字矩形と、矩形間の距離を示す。図19は、矩形間隔のヒストグラムを示す。

【0042】このヒストグラムにおいて、最も距離の短いピークは、漢字の偏とつくりの間隔や、プロポーショナル英字の同一単語内の文字間距離に現れる傾向がある。これらを統合しても異なる文字種がブロックに入ることば少ないので、それらを統合することでブロックデータを形成する。この処理を行うことによってプロポーショナルの単語や一字が分離する(つまり偏とつくりからなる)漢字が一つに統合されることになる。

【0043】また、最も距離の長いピークは、単語間の距離、句読点と次の文字との距離に現れることが多い。これらは(特に単語間の距離は)文字種が変わる場合の境目に用いられることが多く、同一ブロックになることを避けたい。そこで、最も距離の長いピーク値以上の距離の文字矩形については、同一ブロックにしないように処理する。

【0044】さらに、対象矩形の両隣の矩形との距離(A, B)を測定し、その差(A-B)が所定の閾値以上のとき、長い方の距離の矩形同志は統合せず、短い方の距離の矩形を統合するように処理する。図20は、矩形間の間隔の差が大きい位置で矩形の統合を行わない場合を説明する図である。図20では、差が所定の閾値以上大きい位置で矩形の統合を行わないので、3つのブロックが形成される。このような処理を行うことによって、プロポーショナルの英文などで、単語間の距離が絶対的に近くても、文字間距離とは差があるはずであるので、一つの単語だけをまとめて統合できる。また、プロポーショナルフォントであっても日本語の漢字部分は比較的等間隔に配置されるので、日本語文をまとめる場合にも都合がよい。

【0045】上記したブロック抽出方法を用いることによって、英文の場合、日本語文書と違って単語と単語の間は半角相当のスペースで区切られるために、他の文字種と混合してブロックデータとなることが避けられる。

【0046】続いて、ブロック内文字種判別部904は、ブロック毎の日英判別を行う(ステップ1004)。これも前掲した出願の方法を用いればよい。つまり、ブロック内文字種判別部904は、上記処理によってブロック化されたまとまりが、日本語であるか、英数字であるかという文字種の判定を行う。ブロック内は同一文字種として判断する。この文字種の判定は次のように行う。すなわち、ブロック内の矩形の幅に対して、該矩形の垂直方向の黒ランの数または白黒反転回数が所定の閾値以上のとき日本語文字と識別し、抽出されたブロック内の矩形の垂直方向座標値を基に英字を識別する。図21(a)、(b)は、日本語と英字の場合の垂直方向ランの数の具体例を示す。英数字ではノイズがない理想的な場合、最大で“g”の文字で4つのランができる(図21(b))。従って、5つ以上のランがカウントされる場合は日本語とする。図21(a)に示す文字「像」の場合、垂直方向のランの数は、文字の下で示すように変化する。

【0047】日英判別手段905は、ブロック毎の判別結果を集計して当該領域の日英判別を行う(ステップ1005)。ここで、日本語と判定されたブロックの数をJCNT、英語と判定されたブロックの数をECNT、不定と判定されたブロックの数をNCNTとする。図11は、ステップ1005の詳細のフローチャートである。JCNT*Th3 > ECNTのときは日本語と判定し(ステップ1101、1105)、そうではなく、ECNT > JCNTのときは英語と判定する(1102、1104)。それ以外の場合はリジェクトとする(ステップ1103)。しきし値Th3は、例えば2とする。

【0048】〈実施例5〉上記した実施例4では、文字領域単位で日英の判定を行っている。この場合、文字領域によっては文字数が非常に少ない場合がある。そのよ

うな場合は、矩形の数が十分に得られないのでブロックの判別結果数の比率で日英判定を行うことが難しくなる可能性がある。実施例5は、ブロックの数が十分でない場合の実施例である。

【0049】図12は、実施例5の処理フローチャートを示す。日英判別手段105は、集計された文字領域内のブロックの数が十分であるか否か（つまり所定の閾値Th以上あるか否か）を調べ（ステップ1201）、十分でない場合には、前掲した特開平6-150061号公報に記載されているOCRを利用した日英判別を行う（ステップ1203）。この場合は、文字の数が少ないのでOCR処理を施しても処理時間の増大は少なくてすむ。そして、ブロックの数が十分である場合には実施例4で説明したブロック毎の判別結果による日英の識別を行う（ステップ1202）。

【0050】〈実施例6〉実施例6は、実施例4の文字領域毎の日英判別を、ページ単位の日英判別に変更したものである。実施例6の処理フローチャートは、図6、7を用いる。

【0051】図6の処理においては、JCNT、ECNT、NCNTの集計を文字領域毎でなくページ全体について行い、その結果を使用して、前述した図11の処理方法によって日英の判定を行う。このときTh3は文字領域単位の場合とは異なってもよい。

【0052】図7の処理においては、まず、各文字領域毎に判別し、その結果から当該ページの日英判定を行う。具体的には、日本語領域と判定された領域の数をJn、英語領域と判定された領域の数をEnとして、Jn>Enなら日本語ページ、En>Jnなら英語ページと判定する。Jn=Enの場合はリジェクトとしてもよいし、日英の何れかにしてもよい。

【0053】〈実施例7〉実施例7では、文字領域毎またはページ単位で日英判別を行う際に、図13に示すように矩形長を利用する日英判別処理（ステップ1301）と、ブロック毎の判別結果を利用する日英判別処理（ステップ1302）によって、それぞれ日英の判別を行う。そして、それぞれの判別結果から最終的に日英に判別を行う（ステップ1303）。

【0054】両者共に日本語または英語と判定された場合には、最終結果はそのまま日本語または英語と判定すればよい。何れかがリジェクトと判定された場合には、リジェクトでない方の判定結果を最終結果とする。

【0055】両者の判定結果が、一方が日本語で、他方が英語で、その結果が一致しない場合には、以下のいずれかの判定をする。

(1) リジェクトとする。

(2) 両者の確信度を算出し、値の大きな方の結果を採用する。

矩形長を利用する判別方法の確信度としては、例えば

$LCNT / (NCNT + SCNT) > Th1$ で、Th1 50

$= 0.3$ の場合には $LCNT / (NCNT + SCNT) * 2.5$ の値（ただし上限を1とする）

$NCNT / (LCNT + SCNT) < Th2$ で、Th2 = 3 の場合には $(LCNT + SCNT) / NCNT * 2.5$ の値（ただし上限を1とする）

$NCNT / (LCNT + SCNT) > Th2$ で、Th2 = 3 の場合には $NCNT / (LCNT + SCNT) * 0.33$ の値（ただし上限を1とする）とする。

10 【0056】ブロック毎の判別結果を利用する判別方法の確信度としては、例えば

$JCNT * Th3 > ECNT$ で、Th3 = 2 の場合には、 $JCNT / (ECNT * 3)$ の値（ただし上限を1とする）

$ECNT > JCNT$ の場合には、 $ECNT / JCNT * 0.7$ の値（ただし上限を1とする）とする。

【0057】〈実施例8〉図14は、実施例8の構成を示す。また、図15は、実施例8の処理フローチャートを示す。この実施例では、入力された文書のページ全体について、日英判別部1412は、前述した実施例3、6の方法を用いて、そのページが日本語であるか英語であるかの日英識別処理を行い（ステップ1501、1502）、その判別結果に基づいて選択部1403は英文文書認識部1404または日本語文書認識部1405を選択し、選択された言語の文書認識処理を行い（ステップ1504、1505）、その認識結果をディスプレイなどの出力部に出力する（ステップ1506）。

30 【0058】なお、日本語と英語とはその属性が異なることから、領域分割処理やフォント識別処理なども切り替えた方がよい場合がある。そこで、本実施例の文書認識部は、文字認識処理だけではなく、上記した領域分割処理やフォント識別処理も含まれている。

【0059】〈実施例9〉図16は、実施例9の構成を示し、図17は、実施例9の処理フローチャートを示す。実施例8と異なる点は、日英識別を文字領域毎に行う点である。そのために、領域分割部1602は、入力文書を文字領域に分割する（ステップ1701、1702）。ここで、領域分割部では、日英両方に適応できる領域分割方法を使用する。分割処理された後、日英判別部1603は文字領域毎に、例えば前述した実施例1の方法を用いて日英識別処理を行い（ステップ1704）、その判別結果に基づいて選択部1604は英文文書認識部1605または日本語文書認識部1606を選択し、選択された言語の文書認識処理を行い（ステップ1705、1706）、その認識結果をディスプレイなどの出力部1607に出力する（ステップ1707）。

なお、実施例9の文書認識部では、文書認識処理の他にフォント識別処理も行う。

【0060】〈実施例10〉前述した各実施例は、黒画

素連結成分や矩形長を特徴量として日本語と英語を判定している。しかし、黒画素連結成分を用いる判定方法は処理時間がかかり、また矩形長を利用する方法はリジェクトの発生が高くなることもある。なお、外接矩形の上辺、下辺の行内での相対位置の頻度分布のピーク位置を基に和文か英文かを識別する方法もあるが（特公平7-21817号公報を参照）、傾きがある文書が入力された場合には、頻度分布が大きく変化し、識別精度が低下してしまうという問題点がある。

【0061】そこで、本実施例では、行高さに対する、行内の外接矩形の高さのヒストグラムを用いて日本語と英語を識別することにより、文書画像の領域毎に精度よくかつ高速に日本語と英語を識別するものである。そして、上記した日英識別方法でも判別不可能な領域に対しては、別の方法を用いて日英識別を行う。

【0062】図22は、実施例10の構成を示す。また、図23は、実施例10の全体の処理フローチャートである。まず、画像入力手段2201は、文書を読み取ることによって文書画像を得る（ステップ2301）。この画像入力手段は、例えばスキャナ、ファックスなどであり、またデータ通信手段2207を介してネットワーク経由で別の機器から画像を得るようにしてもよい。

【0063】次に、領域生成手段2202は、文字領域を生成する（ステップ2302）。この領域生成方法として、例えば特開平6-20092号公報に記載された方法を用いればよい。次に、行切り出し手段2203は、文字領域から文字認識のための行の切り出しを行なう。つまり、文字の外接矩形を求め、それらを統合して行を生成する（ステップ2303）。日英識別手段2204は、生成した文字領域について日英識別を行なう（ステップ2304）。

【0064】日英の識別は以下のようにして行う。図27は、日英識別（ステップ2304）の詳細のフローチャートである。図24は、切り出された行と行内の外接矩形の一例を示す。まず、行高さに対する、行内の外接矩形高さの割合の頻度分布を算出する（ステップ2701、2702）。行高さを $lineheight$ 、矩形高さを $height$ とする。割合を $height\ rate = height * 100 / lineheight$ とする。また、図25のような傾きのある文書の場合は、より精度良く日英識別するために、行高さの代わりにその行の矩形の高さの最大値を $lineheight$ として用いてもよい。つまり、傾きのある入力文書については、行内矩形の最大高さに対する、行内各外接矩形高さの割合のヒストグラムを基に日英識別する。

【0065】上記した割合 $height\ rate$ が例えば80以上の場合の矩形数を $lcnt$ とし、 $height\ rate$ が例えば70以上80未満の場合の矩形数を $ncnt$ とし、 $height\ rate$ が例えば40以上70未満の場合の矩形数を $scnt$ とする。文字領域内

のすべての矩形に対し、 $lcnt$ 、 $ncnt$ 、 $scnt$ を求める。

【0066】図26は、日本語文書と英語文書について調べた矩形数の一例を示す。一般に、日本語は $lcnt$ が大きく、英語は $scnt$ が大きいという傾向がある。そこで、所定の閾値 thJ 、 thE を設定し、 $lcnt / scnt > thJ$ のとき日本語と判定し（ステップ2703）、 $lcnt / scnt < thE$ のとき英語と判定する（ステップ2704）。それ以外のときは不明領域とする（ステップ2705）。

【0067】上記した不明領域に対して、統計的手法を用いて日英識別することができる。図28は、不明領域に対する詳細な処理フローチャートである。例えば、あらかじめ日本語領域と英語領域の特徴値 $lcnt$ 、 $ncnt$ 、 $scnt$ を正規化し、その平均値と共分散行列の逆行列を日本語、英語についてそれぞれ求める。そして、平均値と共分散行列の逆行列を用いて、日本語、英語のそれぞれについてマハラノビス距離を求める（ステップ2801、2802）。

【0068】日本語のマハラノビス距離を Dj 、英語のマハラノビス距離を De とすると、所定の閾値を Me 、 Mj とすると、 $Dj / De > Me$ のとき英語と判定し（ステップ2803）、 $Dj / De < MJ$ のとき日本語と判定する（ステップ2804）。何れの条件にも満足しない場合は不明領域と判定する（ステップ2805）。なお、上記したマハラノビス距離の代わりに、平均値とのユークリッド距離やシティブロック距離を用いてもよい。

【0069】さらに不明と判定された領域に対して、英文認識の確信度を用いて日英識別を行う。図29は、ステップ2805の詳細な処理フローチャートである。英文認識で確信度を算出する（ステップ2901）。次いで、算出された確信度について、例えば60%以上の確信度をもつ単語の個数を $Good$ 、60%未満で確信度0でない単語の個数を Bad 、確信度が0の単語の個数を $Zelo$ とする（ステップ2902）。

【0070】日英識別の判定値を $Value$ とすると、 $Value = Good / (Good + Bad + Zelo)$

とし（ステップ2903）、 $Value$ が所定の閾値 $th\ eocr$ を超えれば（ステップ2904）、英語と判定し、それ以下ならば日本語と判定する。

【0071】なお、 $Zelo$ に重み付けしてもよい。 $Zelo$ を例えば Bad の3個分とすると、 $Value$ は、 $Bad = Bad + Zelo \times 3$ であるから
 $Value = Good / (Good + Bad)$

となり、 $Value$ が閾値 $th\ eocr$ を超えれば英語、それ以下ならば日本語と判定することもできる。このように、日英識別判定のための文字数が少ない領域でも、英文認識による確信度で日英識別しているので、精

度よく領域単位の日英識別が行われる。

【0072】〈実施例11〉本実施例は、入力文書画像を縮小した画像から外接矩形を生成し、生成された矩形同士で適当な統合を行い、統合後の矩形長の縦横比のヒストグラムを用いて日英識別をより精度良く行なう実施例である。

【0073】図30は、実施例11の構成を示す。また、図31は、実施例11の全体の処理フローチャートである。上記した実施例と同様にして画像入力手段3001によって入力された文書画像は、画像縮小手段3002によって縮小される（ステップ3101、3102）。この処理は、例えば文書画像を1/4程度にOR圧縮（4×4画素を1画素に縮小し、16画素中に1つでも黒画素があれば縮小画像は黒とする）する。

【0074】次に、領域生成手段3003は、文字領域を生成する（ステップ3103）。この領域生成方法として、例えば特開平6-20092号公報に記載された方法を用いればよい。続いて、矩形統合手段3004は、日英の特性が良く表れるように、矩形の統合を行なう（ステップ3104）。例えば、図32に示すように、矩形1、2のy座標（縦方向）の上下座標が近くかつ、隣同士の矩形1、2のx座標が非常に近い場合（例えば、矩形間の水平距離が英語のスペースに相当する距離より小さい場合）、矩形を統合する。また、例えば、図33に示すように、左側の矩形1が右側の矩形2をy座標で包含する位置関係にありかつ、隣同士の矩形1、2のx座標が非常に近い場合（例えば、矩形間の水平距離が英語のスペースに相当する距離より小さい場合）、矩形を統合する。

【0075】そして、矩形縦横比（矩形長縦／矩形長横）を用いて、長矩形、中矩形、小矩形、極小矩形の4つの特徴量に分ける（図34）。一般に、日本語は長矩形の出現する割合が高く、また、英語は中矩形の出現する割合が高い。この特性の違いを利用して、日英識別手段3005は、識別判定式を作成し、日英識別を行なう（ステップ3105）。図35は、日英識別処理の詳細のフローチャートである。

【0076】

例えば、領域内での長矩形の領域数 $lcnt$

領域内での中矩形の領域数 $ncnt$

領域内での小矩形の領域数 $scnt$

領域内での極小矩形の領域数 $sscnt$ （ノイズの場合が多い）を算出し（ステップ3501）、領域内での長矩形の割合 $ratio1 = lcnt / (ncnt + scnt)$ を算出し（ステップ3502）、領域内での中矩形の割合 $ratio2 = ncnt / (lcnt + scnt)$ を算出する（ステップ3503）。なお、上記割合を算出するとき、 $sscnt$ はノイズとして無視した。

【0077】そして、 $ratio1$ をx座標、 $ratio2$ をy座標とし、誤識別を極力少なく、日英重なって

いる部分はリジェクトになるように、日本語領域、英語領域、リジェクト領域に分ける。例えば、 $ratio2 / ratio1 > thE$ ならば英語領域と判定（ステップ3504）し、 $ratio2 / ratio1 < thJ$ ならば日本語領域と判定し（ステップ3505）、それ以外の領域は日英不明とする（ステップ3506）。ここで、 thE 、 thJ は所定の閾値である。

【0078】日英不明と判定された領域に対して、実施例10と同様に、統計的手法を用いて日英識別する。例えば、あらかじめ日本語領域と英語領域の特徴値 $lcnt$ 、 $ncnt$ 、 $scnt$ を正規化し、その平均値と共分散行列の逆行列を日本語、英語でそれぞれ求める。平均値と共分散行列の逆行列を用いて日本語、英語のそれぞれのマハラノビス距離を求める。日本語のマハラノビス距離を Dj 、英語のマハラノビス距離を De とするとき、所定の閾値を Me 、 Mj とすると、 $Dj / De > Me$ のとき英語、 $Dj / De < MJ$ のとき日本語と判定する。何れの条件も満たさない場合は不明と判定する。なお、マハラノビス距離の代わりに、平均値とのユークリッド距離やシティブロック距離を用いてもよい。

【0079】〈実施例12〉本発明は上記した実施例に限定されず、ソフトウェアによっても実現することができる。本発明をソフトウェアによって実現する場合には、図36に示すように、CPU、メモリ、表示装置、ハードディスク、キーボード、CD-ROMドライブ、スキャナなどからなるコンピュータシステムを用意し、CD-ROMなどのコンピュータ読み取り可能な記録媒体には、本発明の日本語英語判定機能、文書認識機能を実現するプログラムなどが記録されている。また、スキャナなどの画像入力手段から入力された文書画像などは一時的にハードディスクなどに格納される。そして、該プログラムが起動されると、一時保存された文書画像データが読み込まれて、日本語英語判定処理、文書認識処理を実行し、その結果をディスプレイなどに出力する。

【0080】

【発明の効果】以上、説明したように、請求項1、12記載の発明によれば、複数の判定方法を併用しているので、高精度に日本語と英語とを判別することができる。

【0081】請求項2、3、6、7、13、14記載の発明によれば、文書画像中の文字領域毎に精度よく日本語と英語の判別を行うことができる。

【0082】請求項4、5、8、9、13、14記載の発明によれば、文書画像のページ単位に、精度よく日本語と英語の判別を行うことができる。

【0083】請求項10、11、15記載の発明によれば、日本語または英語と判定された文書画像に対して、適切な文書認識処理を実行しているので、高精度な認識結果を得ることができる。

【図面の簡単な説明】

【図1】本発明の実施例1の構成を示す。

【図 2】本発明の実施例 1 の全体の処理フローチャートを示す。

【図 3】英文、日本語文の画像例と、その外接矩形を示す。

【図 4】実施例 1 の日英判定の処理フローチャートを示す。

【図 5】実施例 2 の処理フローチャートを示す。

【図 6】実施例 3 に係るステップ 205 の第 1 の詳細フローチャートを示す。

【図 7】実施例 3 に係るステップ 205 の第 2 の詳細フローチャートを示す。

【図 8】ステップ 403 の詳細フローチャートを示す。

【図 9】実施例 4 の構成を示す。

【図 10】実施例 4 の処理フローチャートを示す。

【図 11】ステップ 1005 の詳細のフローチャートである。

【図 12】実施例 5 の処理フローチャートを示す。

【図 13】実施例 7 の処理フローチャートを示す。

【図 14】実施例 8 の構成を示す。

【図 15】実施例 8 の処理フローチャートを示す。

【図 16】実施例 9 の構成を示す。

【図 17】実施例 9 の処理フローチャートを示す。

【図 18】抽出された文字矩形と、矩形間の距離を示す。

【図 19】矩形間隔のヒストグラムを示す。

【図 20】矩形間の間隔の差が大きい位置で矩形の統合を行わない場合を説明する図である。

【図 21】(a)、(b) は、日本語と英字の場合の垂直方向ランの数の具体例を示す。

【図 22】実施例 10 の構成を示す。

【図 23】実施例 10 の全体の処理フローチャートである。

*【図 24】切り出された行と行内の外接矩形の一例を示す。

【図 25】文書が傾いている場合の行と行内の外接矩形の一例を示す。

【図 26】日本語文書と英語文書について調べた矩形数の一例を示す。

【図 27】日英識別（ステップ 2304）の詳細な処理フローチャートである。

【図 28】不明領域に対する詳細な処理フローチャートである。

【図 29】ステップ 2805 の詳細な処理フローチャートである。

【図 30】実施例 11 の構成を示す。

【図 31】実施例 11 の全体の処理フローチャートである。

【図 32】矩形を統合する例を示す。

【図 33】矩形を統合する他の例を示す。

【図 34】4 種類に分類された矩形を示す。

【図 35】実施例 11 の日英識別処理の詳細な処理フローチャートである。

【図 36】実施例 12 の構成を示す。

【符号の説明】

101 画像入力手段

102 画像縮小手段

103 連結成分抽出手段

104 領域生成手段

105 日英判別手段

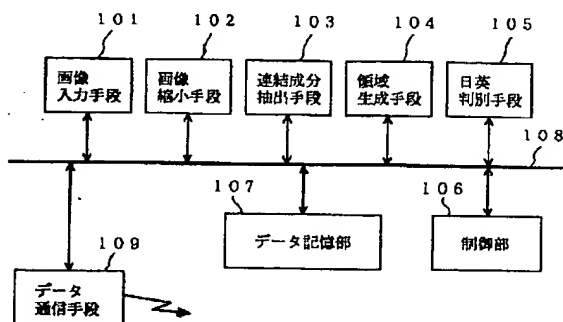
106 制御部

107 データ記憶部

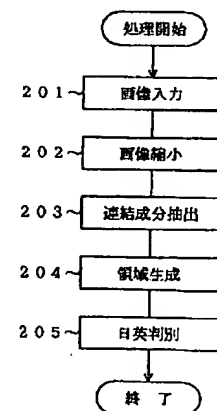
108 データ通信路

109 データ通信手段

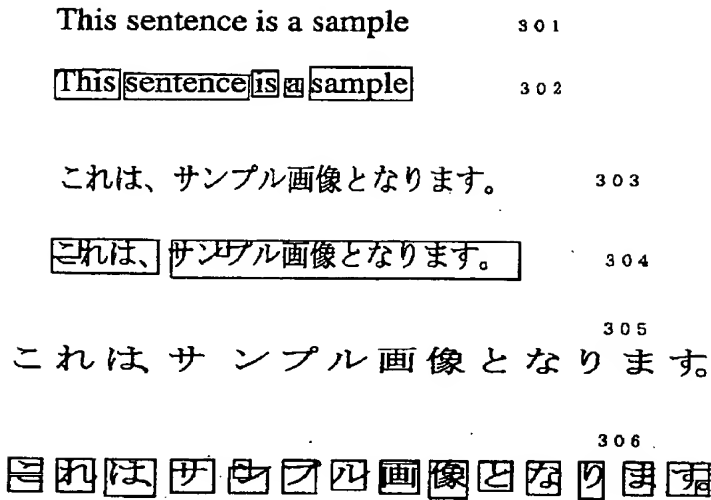
【図 1】



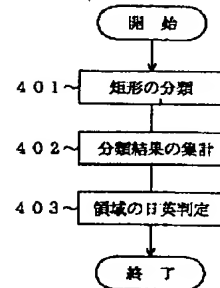
【図 2】



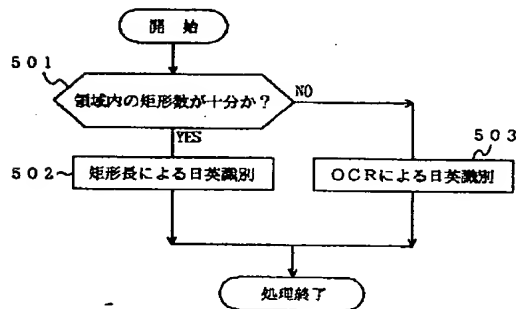
【図3】



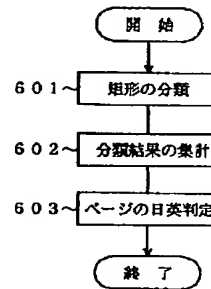
【図4】



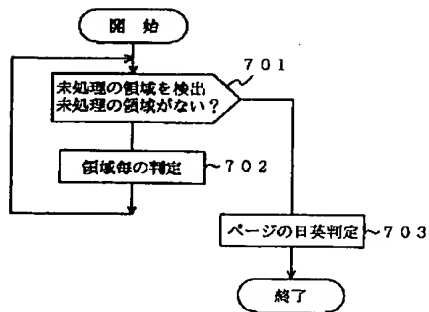
【図5】



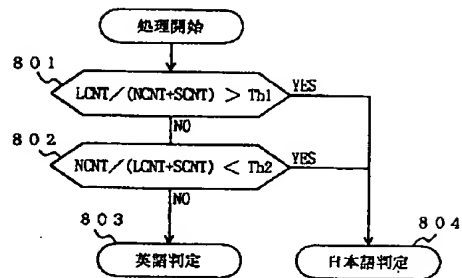
【図6】



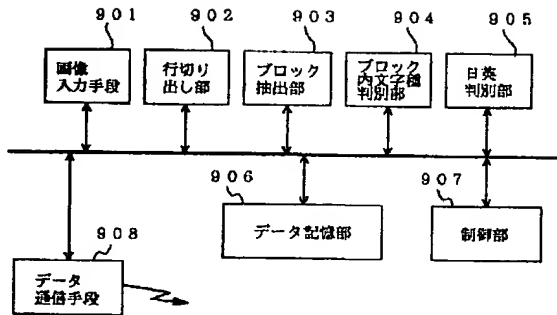
【図7】



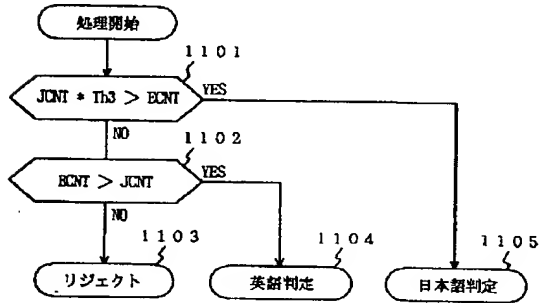
【図8】



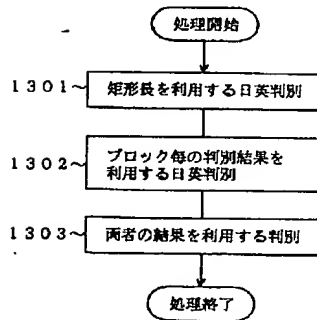
【図9】



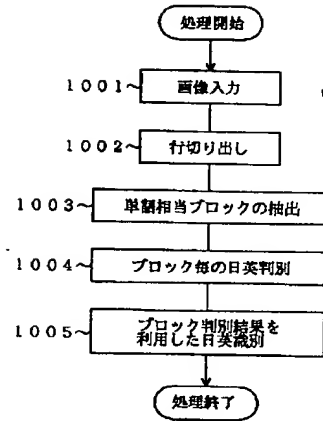
【図11】



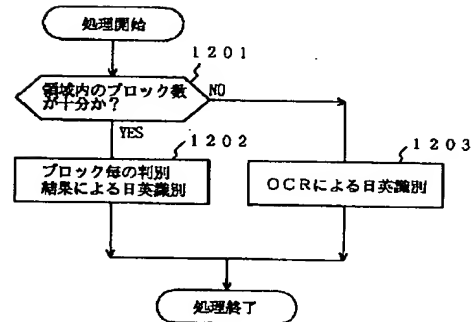
【図13】



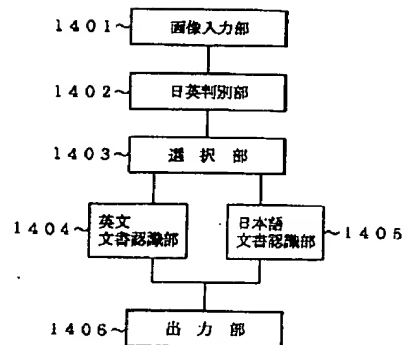
【図10】



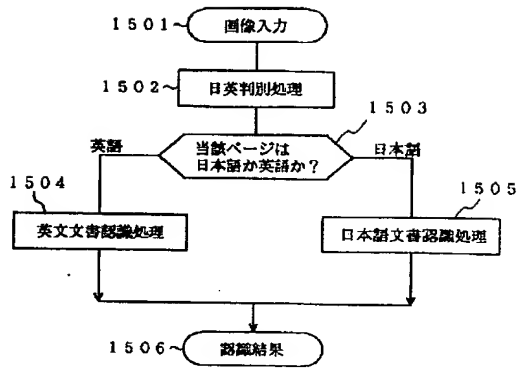
【図12】



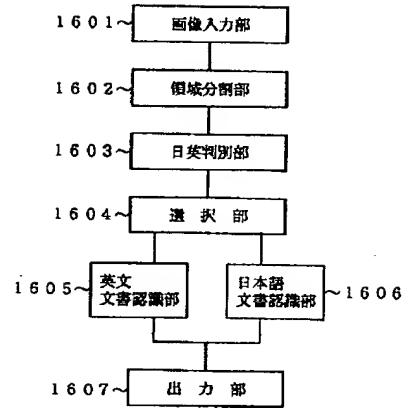
【図14】



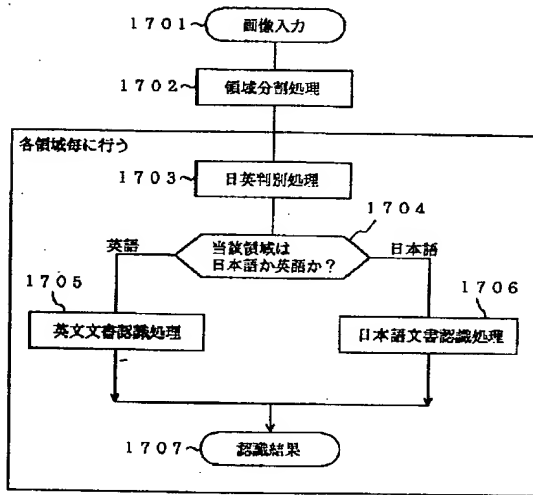
【図15】



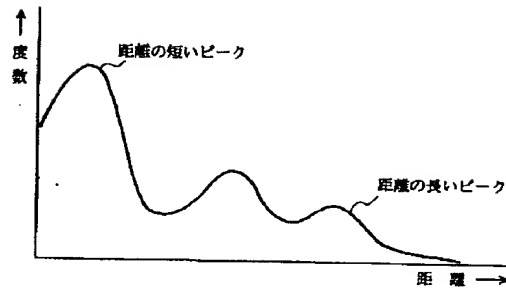
【図16】



【図17】

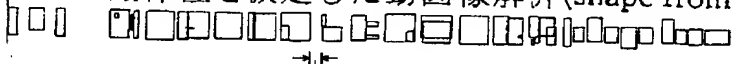


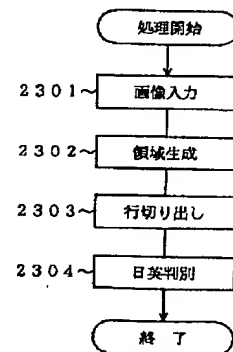
【図19】



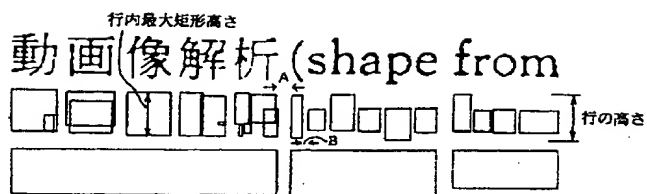
【図23】

【図18】

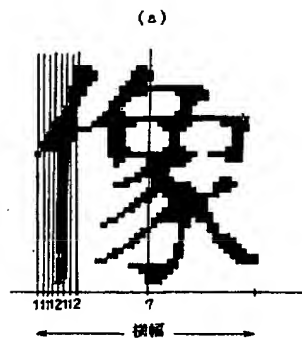
(4) 剛体性を仮定した動画像解析(shape from

 矩形間距離



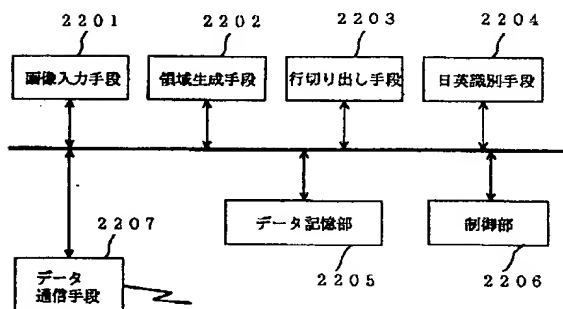
【図20】



【図21】



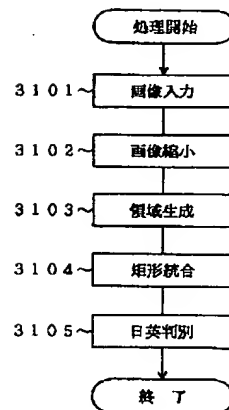
【図22】



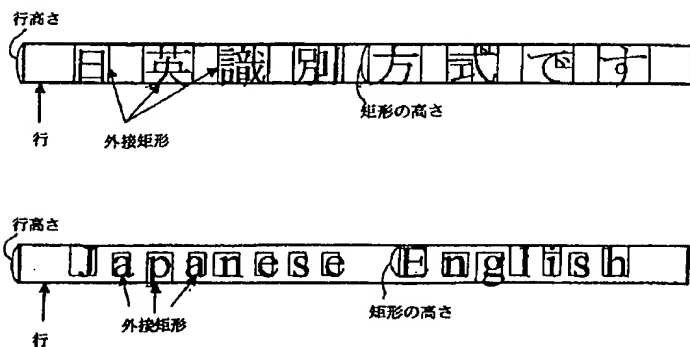
(a)



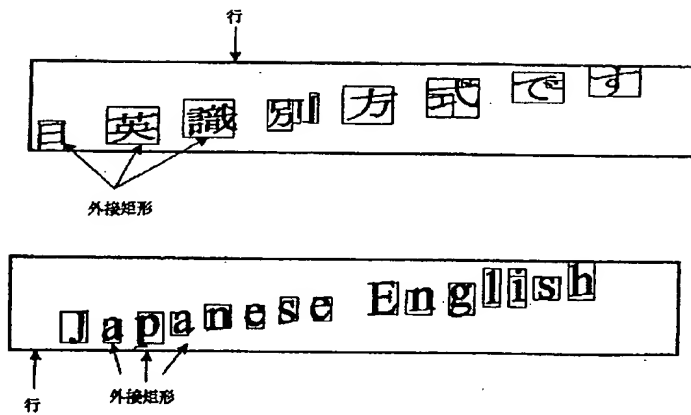
【図31】



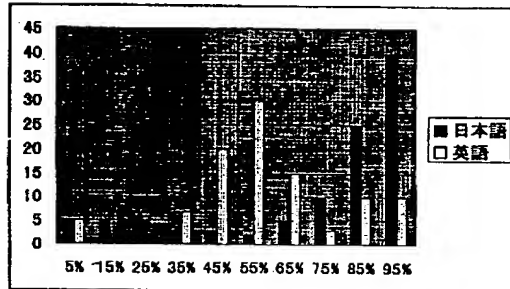
【図24】



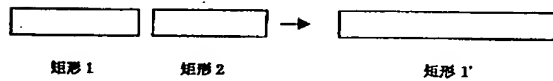
【図25】



【図26】



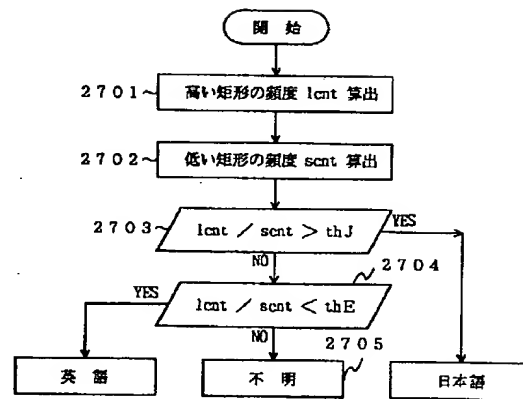
【図32】



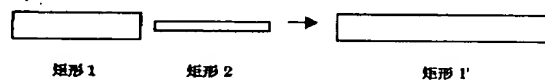
【図34】

長矩形	縦横比 7 以上 (領域長が短い場合 5 以上)
中矩形	縦横比 1.8 以上
幅短矩形	横の長さが 16 画素以下
短矩形	それ以外

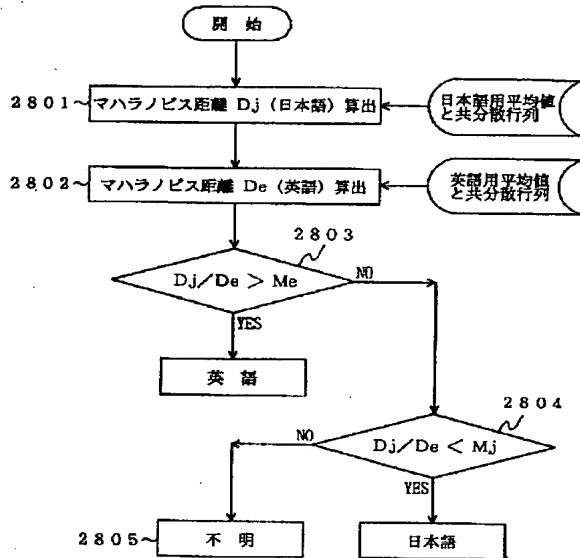
【図27】



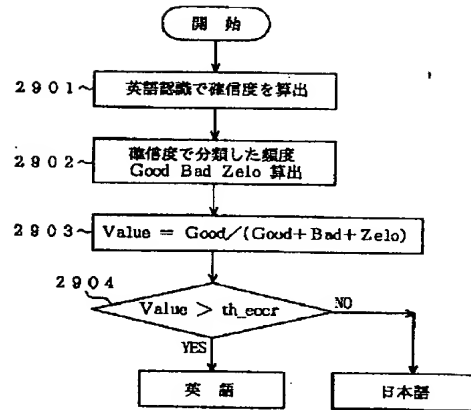
【図33】



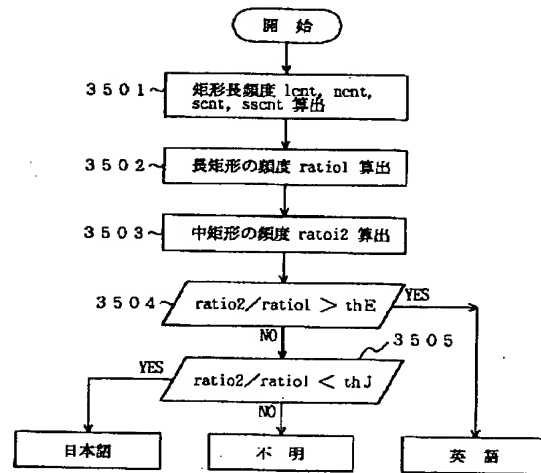
【図28】



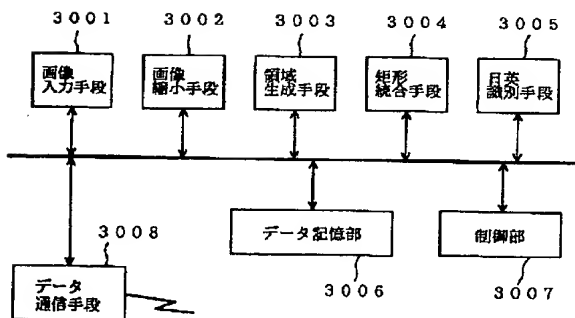
【図29】



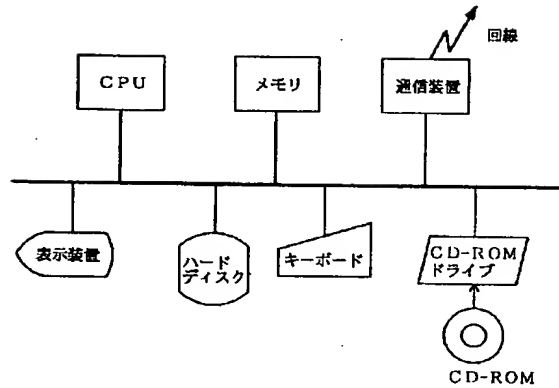
【図35】



【図30】



【図36】



【手続補正書】

【提出日】平成10年7月15日

【手続補正1】

【補正対象書類名】明細書

【補正対象項目名】全文

【補正方法】変更

【補正内容】

【書類名】明細書

【発明の名称】文書画像の日本語英語判定方法、文書認識方法および記録媒体

【特許請求の範囲】

【請求項1】 文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、複数の判定方法を用いて日本語領域であるか英語領域であるかを判定し、該複数の判定結果を比較することによって最終判定結果を得ることを特徴とする文書画像の日本語英語判定方法。

【請求項2】 文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、前記文書画像を縮小することにより生成される文字領域内の黒画素連結成分の長さを基に該連結成分を分類し、該分類結果の集計値を基に前記各文字領域が日本語領域であるか英語領域であるかを判定することを特徴とする文書画像の日本語英語判定方法。

【請求項3】 前記生成される文字領域内の黒画素連結成分の数が所定の条件を満たさないとき、異なる判定方法を用いることを特徴とする請求項2記載の文書画像の日本語英語判定方法。

【請求項4】 各ページの文書画像が日本語文書画像であるか英語文書画像であるかを判定する文書画像の日本

語英語判定方法であって、前記文書画像を縮小することにより生成されるページ内の黒画素連結成分の長さを基に該連結成分を分類し、該分類結果の集計値を基に前記各ページが日本語領域であるか英語領域であるかを判定することを特徴とする文書画像の日本語英語判定方法。

【請求項5】 ページが複数の文字領域からなり、各ページの文書画像が日本語文書画像であるか英語文書画像であるかを判定する文書画像の日本語英語判定方法であって、前記文書画像を縮小することにより生成される文字領域内の黒画素連結成分の長さを基に該連結成分を分類し、該分類結果の集計値を基に前記各文字領域が日本語領域であるか英語領域であるかを判定し、該判定結果を基に前記各ページが日本語領域であるか英語領域であるかを判定することを特徴とする文書画像の日本語英語判定方法。

【請求項6】 文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、前記文字領域中から行を検出し、該行中から近接した外接矩形を統合してブロックを抽出し、該ブロック毎に日本語領域であるか英語領域であるか、あるいは判定不能領域であるかを判定し、該判定結果を前記ブロック毎に集計し、該集計値を基に前記各文字領域が日本語領域であるか英語領域であるかを判定することを特徴とする文書画像の日本語英語判定方法。

【請求項7】 前記抽出されるブロックの数が所定の条件を満たさないとき、異なる判定方法を用いることを特徴とする請求項6記載の文書画像の日本語英語判定方法。

【請求項8】 ページが複数の文字領域からなり、各ペ

ページの文書画像が日本語文書画像であるか英語文書画像であるかを判定する文書画像の日本語英語判定方法であって、前記文字領域中から行を検出し、該行中から近接した外接矩形を統合してブロックを抽出し、該ブロック毎に日本語領域であるか英語領域であるか、あるいは判定不能領域であるかを判定し、該判定結果をページ単位で集計し、該集計値を基に前記各ページが日本語文書画像であるか英語文書画像であるかを判定することを特徴とする文書画像の日本語英語判定方法。

【請求項 9】 ページが複数の文字領域からなり、各ページの文書画像が日本語文書画像であるか英語文書画像であるかを判定する文書画像の日本語英語判定方法であって、前記文字領域中から行を検出し、該行中から近接した外接矩形を統合してブロックを抽出し、該ブロック毎に日本語領域であるか英語領域であるか、あるいは判定不能領域であるかを判定し、該判定結果を文字領域毎に集計し、該集計値を基に文字領域毎に日本語領域であるか英語領域であるかを判定し、該判定結果をページ単位で集計し、該集計値を基に前記各ページが日本語文書画像であるか英語文書画像であるかを判定することを特徴とする文書画像の日本語英語判定方法。

【請求項 10】 文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、前記各文字領域から行を切り出し、行の高さと行内の矩形高さを基に前記各文字領域が日本語領域であるか英語領域であるかを判定することを特徴とする文書画像の日本語英語判定方法。

【請求項 11】 前記行の高さと行内の矩形高さは、行の高さに対する行内の各矩形高さの割合のヒストグラムであることを特徴とする請求項 10 記載の文書画像の日本語英語判定方法。

【請求項 12】 文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、前記各文字領域から行を切り出し、行の高さに対する行内の各矩形高さの割合のヒストグラムを基に前記各文字領域が日本語領域であるか英語領域であるかを判定し、何れの領域にも判定できない不明領域については、予め前記ヒストグラムを基に日本語の特性値と英語の特性値とを算出しておき、前記不明領域が前記何れの特性値に近いかを算出することによって日本語領域であるか英語領域であるかを判定することを特徴とする文書画像の日本語英語判定方法。

【請求項 13】 文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、前記各文字領域から行を切り出し、行内の矩形の最大高さと行内の矩形高さを基に前記各文字領域が日本語領域であるか英語領域であるかを判定することを特徴とする文書画像の日本語英語判定方法。

【請求項 14】 前記行内の矩形の最大高さと行内の矩

形高さは、行内の矩形の最大高さに対する行内の各矩形高さの割合のヒストグラムであることを特徴とする請求項 13 記載の文書画像の日本語英語判定方法。

【請求項 15】 文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、前記各文字領域から行を切り出し、行内の矩形の最大高さに対する行内の各矩形高さの割合のヒストグラムを基に前記各文字領域が日本語領域であるか英語領域であるかを判定し、何れの領域にも判定できない不明領域については、予め前記ヒストグラムを基に日本語の特性値と英語の特性値とを算出しておき、前記不明領域が前記何れの特性値に近いかを算出することによって日本語領域であるか英語領域であるかを判定することを特徴とする文書画像の日本語英語判定方法。

【請求項 16】 前記 2 度目の判定でも判定できない不明領域については、英文認識の確信度を基に日本語領域であるか英語領域であるかを判定することを特徴とする請求項 12 または 15 記載の文書画像の日本語英語判定方法。

【請求項 17】 文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、前記文書画像を縮小した画像から外接矩形を生成し、所定の位置関係にある矩形同士を統合し、統合後の矩形の縦横比のヒストグラムを基に前記各文字領域が日本語領域であるか英語領域であるかを判定することを特徴とする文書画像の日本語英語判定方法。

【請求項 18】 文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、前記文書画像を縮小した画像から外接矩形を生成し、所定の位置関係にある矩形同士を統合し、統合後の矩形の縦横比のヒストグラムを基に前記各文字領域が日本語領域であるか英語領域であるかを判定し、何れの領域にも判定できない不明領域については、予め前記ヒストグラムを基に日本語の特性値と英語の特性値とを算出しておき、前記不明領域が前記何れの特性値に近いかを算出することによって日本語領域であるか英語領域であるかを判定することを特徴とする文書画像の日本語英語判定方法。

【請求項 19】 文書画像が日本語文書画像であるか英語文書画像であるかを判定し、該判定結果に応じた文書認識処理を行うことを特徴とする文書認識方法。

【請求項 20】 文書画像を複数の文字領域に分割し、該分割された文字領域毎に日本語文書領域であるか英語文書領域であるかを判定し、該判定結果に応じた文書認識処理を行うことを特徴とする文書認識方法。

【請求項 21】 文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定するために、複数の判定方法を用いて日本語領域であるか英語領域であるかを

判定する機能と、該複数の判定結果を比較することによって最終判定結果を得る機能をコンピュータに実現させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項22】 文書画像中の各文字領域または各ページの文書画像が日本語領域であるか英語領域であるかを判定するために、前記文書画像を縮小することにより生成される文字領域内またはページ内の黒画素連結成分の長さを基に該連結成分を分類する機能と、該分類結果の集計値を基に前記各文字領域または各ページが日本語領域であるか英語領域であるかを判定する機能をコンピュータに実現させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項23】 文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定するために、または、ページが複数の文字領域からなり、各ページの文書画像が日本語文書画像であるか英語文書画像であるかを判定するために、前記文字領域中から行を検出する機能と、該行中から近接した外接矩形を統合してブロックを抽出する機能と、該ブロック毎に日本語領域であるか英語領域であるか、あるいは判定不能領域であるかを判定する機能と、該判定結果を前記ブロック毎またはページ単位に集計する機能と、該集計値を基に、前記各文字領域が日本語領域であるか英語領域であるかを判定する機能または各ページが日本語文書画像であるか英語文書画像であるかを判定する機能をコンピュータに実現させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項24】 文書画像が日本語文書画像であるか英語文書画像であるかを判定する機能または文書画像を複数の文字領域に分割し、該分割された文字領域毎に日本語文書領域であるか英語文書領域であるかを判定する機能と、該判定結果に応じた文書認識処理を行う機能をコンピュータに実現させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項25】 文書画像中の各文字領域から行を切り出す機能と、行の高さに対する行内の各矩形高さの割合のヒストグラムまたは行内の矩形の最大高さに対する行内の各矩形高さの割合のヒストグラムを基に前記各文字領域が日本語領域であるか英語領域であるかを判定する機能と、何れの領域にも判定できない不明領域については、予め前記ヒストグラムを基に日本語の特性値と英語の特性値とを算出する機能と、前記不明領域が前記何れの特性値に近いかを算出することによって日本語領域であるか英語領域であるかを判定する機能をコンピュータに実現させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項26】 文書画像から縮小画像を生成する機能と、該縮小画像から文字領域を生成する機能と、該文字領域から外接矩形を生成する機能と、所定の位置関係に

ある矩形同士を統合する機能と、統合後の矩形の縦横比のヒストグラムを基に前記文字領域が日本語領域であるか英語領域であるかを判定する機能をコンピュータに実現させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、文書画像中の各文字領域に対して日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法および記録媒体に関し、また文書画像が日本語文書画像であるか英語文書画像であるかを判定してから認識処理する文書認識方法および記録媒体に関する。

【0002】

【従来の技術】文書画像に対して文字認識処理を施す場合に、適切な言語を選択する必要がある。すなわち、英文OCRで日本語を認識しようとしてもアルファベットや数字以外は認識不可能であるし、また逆に日本語OCRで英文を認識しようすると、文字切り出しや言語処理のうえで英文OCRを使用した場合よりも認識率が低くなってしまう。

【0003】従って、文字認識処理を施す前に、言語識別を行う必要が生じる。従来から文書中の文字種を識別する種々の手法が提案されている。例えば、2値化された文字行の縦方向または横方向の黒白反転回数を計数し、その分布を基に文字種の識別を行う文書認識装置がある（特開平5-108876号公報を参照）。

【0004】また、読み取った単語を認識させ、その認識結果と辞書との適合率を基に認識文字の言語種類を判別する文書認識装置もある（特開平6-150061号公報を参照）。

【0005】

【発明が解決しようとする課題】上記した前者の装置では、文字種を識別する特徴として黒白反転回数を用いているが、この特徴はフォントや文書内容（かな、漢字、数字などの比率）による変動が大きく、このために識別の精度が低くなるという問題がある。

【0006】これに対して、後者の装置では、一度、文字認識を行っているので、OCRの性能がよければかなりの確率で字種が判明することになり、精度よく日英判別を行うことが可能となる。しかし、OCRは処理に多くの時間を要するという問題がある。

【0007】本発明は上記した事情を考慮してなされたもので、本発明の目的は、精度よくかつ高速に日本語と英語の識別を行うと共に、識別する範囲についても各文字領域毎に、またページ単位毎に両者を識別できる文書画像の日本語英語判別方法および記録媒体、さらには、文書画像を判定し、最適な文書認識処理を行う文書認識方法および記録媒体を提供することにある。

【0008】

【課題を解決するための手段】前記目的を達成するために、請求項1記載の発明では、文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、複数の判定方法を用いて日本語領域であるか英語領域であるかを判定し、該複数の判定結果を比較することによって最終判定結果を得ることを特徴としている。

【0009】請求項2記載の発明では、文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、前記文書画像を縮小することにより生成される文字領域内の黒画素連結成分の長さを基に該連結成分を分類し、該分類結果の集計値を基に前記各文字領域が日本語領域であるか英語領域であるかを判定することを特徴としている。

【0010】請求項3記載の発明では、前記生成される文字領域内の黒画素連結成分の数が所定の条件を満たさないとき、異なる判定方法を用いることを特徴としている。

【0011】請求項4記載の発明では、各ページの文書画像が日本語文書画像であるか英語文書画像であるかを判定する文書画像の日本語英語判定方法であって、前記文書画像を縮小することにより生成されるページ内の黒画素連結成分の長さを基に該連結成分を分類し、該分類結果の集計値を基に前記各ページが日本語領域であるか英語領域であるかを判定することを特徴としている。

【0012】請求項5記載の発明では、ページが複数の文字領域からなり、各ページの文書画像が日本語文書画像であるか英語文書画像であるかを判定する文書画像の日本語英語判定方法であって、前記文書画像を縮小することにより生成される文字領域内の黒画素連結成分の長さを基に該連結成分を分類し、該分類結果の集計値を基に前記各文字領域が日本語領域であるか英語領域であるかを判定し、該判定結果を基に前記各ページが日本語領域であるか英語領域であるかを判定することを特徴としている。

【0013】請求項6記載の発明では、文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、前記文字領域中から行を検出し、該行中から近接した外接矩形を統合してブロックを抽出し、該ブロック毎に日本語領域であるか英語領域であるか、あるいは判定不能領域であるかを判定し、該判定結果を前記ブロック毎に集計し、該集計値を基に前記各文字領域が日本語領域であるか英語領域であるかを判定することを特徴としている。

【0014】請求項7記載の発明では、前記抽出されるブロックの数が所定の条件を満たさないとき、異なる判定方法を用いることを特徴としている。

【0015】請求項8記載の発明では、ページが複数の文字領域からなり、各ページの文書画像が日本語文書画像であるか英語文書画像であるかを判定する文書画像の

日本語英語判定方法であって、前記文字領域中から行を検出し、該行中から近接した外接矩形を統合してブロックを抽出し、該ブロック毎に日本語領域であるか英語領域であるか、あるいは判定不能領域であるかを判定し、該判定結果をページ単位で集計し、該集計値を基に前記各ページが日本語文書画像であるか英語文書画像であるかを判定することを特徴としている。

【0016】請求項9記載の発明では、ページが複数の文字領域からなり、各ページの文書画像が日本語文書画像であるか英語文書画像であるかを判定する文書画像の日本語英語判定方法であって、前記文字領域中から行を検出し、該行中から近接した外接矩形を統合してブロックを抽出し、該ブロック毎に日本語領域であるか英語領域であるか、あるいは判定不能領域であるかを判定し、該判定結果を文字領域毎に集計し、該集計値を基に文字領域毎に日本語領域であるか英語領域であるかを判定し、該判定結果をページ単位で集計し、該集計値を基に前記各ページが日本語文書画像であるか英語文書画像であるかを判定することを特徴としている。

【0017】請求項10記載の発明では、文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、前記各文字領域から行を切り出し、行の高さと行内の矩形高さを基に前記各文字領域が日本語領域であるか英語領域であるかを判定することを特徴としている。

【0018】請求項11記載の発明では、前記行の高さと行内の矩形高さは、行の高さに対する行内の各矩形高さの割合のヒストグラムであることを特徴としている。

【0019】請求項12記載の発明では、文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、前記各文字領域から行を切り出し、行の高さに対する行内の各矩形高さの割合のヒストグラムを基に前記各文字領域が日本語領域であるか英語領域であるかを判定し、何れの領域にも判定できない不明領域については、予め前記ヒストグラムを基に日本語の特性値と英語の特性値とを算出しておき、前記不明領域が前記何れの特性値に近いかを算出することによって日本語領域であるか英語領域であるかを判定することを特徴としている。

【0020】請求項13記載の発明では、文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、前記各文字領域から行を切り出し、行内の矩形の最大高さで行内の矩形高さを基に前記各文字領域が日本語領域であるか英語領域であるかを判定することを特徴としている。

【0021】請求項14記載の発明では、前記行内の矩形の最大高さで行内の矩形高さは、行内の矩形の最大高さに対する行内の各矩形高さの割合のヒストグラムであることを特徴としている。

【0022】請求項15記載の発明では、文書画像中の

各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、前記各文字領域から行を切り出し、行内の矩形の最大高さに対する行内の各矩形高さの割合のヒストグラムを基に前記各文字領域が日本語領域であるか英語領域であるかを判定し、何れの領域にも判定できない不明領域については、予め前記ヒストグラムを基に日本語の特性値と英語の特性値とを算出しておき、前記不明領域が前記何れの特性値に近いかを算出することによって日本語領域であるか英語領域であるかを判定することを特徴としている。

【0023】請求項16記載の発明では、前記2度目の判定でも判定できない不明領域については、英文認識の確信度を基に日本語領域であるか英語領域であるかを判定することを特徴としている。

【0024】請求項17記載の発明では、文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、前記文書画像を縮小した画像から外接矩形を生成し、所定の位置関係にある矩形同士を統合し、統合後の矩形の縦横比のヒストグラムを基に前記各文字領域が日本語領域であるか英語領域であるかを判定することを特徴としている。

【0025】請求項18記載の発明では、文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定する文書画像の日本語英語判定方法であって、前記文書画像を縮小した画像から外接矩形を生成し、所定の位置関係にある矩形同士を統合し、統合後の矩形の縦横比のヒストグラムを基に前記各文字領域が日本語領域であるか英語領域であるかを判定し、何れの領域にも判定できない不明領域については、予め前記ヒストグラムを基に日本語の特性値と英語の特性値とを算出しておき、前記不明領域が前記何れの特性値に近いかを算出することによって日本語領域であるか英語領域であるかを判定することを特徴としている。

【0026】請求項19記載の発明では、文書画像が日本語文書画像であるか英語文書画像であるかを判定し、該判定結果に応じた文書認識処理を行うことを特徴としている。

【0027】請求項20記載の発明では、文書画像を複数の文字領域に分割し、該分割された文字領域毎に日本語文書領域であるか英語文書領域であるかを判定し、該判定結果に応じた文書認識処理を行うことを特徴としている。

【0028】請求項21記載の発明では、文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定するために、複数の判定方法を用いて日本語領域であるか英語領域であるかを判定する機能と、該複数の判定結果を比較することによって最終判定結果を得る機能をコンピュータに実現させるためのプログラムを記録した

コンピュータ読み取り可能な記録媒体であることを特徴としている。

【0029】請求項22記載の発明では、文書画像中の各文字領域または各ページの文書画像が日本語領域であるか英語領域であるかを判定するために、前記文書画像を縮小することにより生成される文字領域内またはページ内の黒画素連結成分の長さを基に該連結成分を分類する機能と、該分類結果の集計値を基に前記各文字領域または各ページが日本語領域であるか英語領域であるかを判定する機能をコンピュータに実現させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体であることを特徴としている。

【0030】請求項23記載の発明では、文書画像中の各文字領域が日本語領域であるか英語領域であるかを判定するために、または、ページが複数の文字領域からなり、各ページの文書画像が日本語文書画像であるか英語文書画像であるかを判定するために、前記文字領域中から行を検出する機能と、該行中から近接した外接矩形を統合してブロックを抽出する機能と、該ブロック毎に日本語領域であるか英語領域であるか、あるいは判定不能領域であるかを判定する機能と、該判定結果を前記ブロック毎またはページ単位に集計する機能と、該集計値を基に、前記各文字領域が日本語領域であるか英語領域であるかを判定する機能または各ページが日本語文書画像であるか英語文書画像であるかを判定する機能をコンピュータに実現させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体であることを特徴としている。

【0031】請求項24記載の発明では、文書画像が日本語文書画像であるか英語文書画像であるかを判定する機能または文書画像を複数の文字領域に分割し、該分割された文字領域毎に日本語文書領域であるか英語文書領域であるかを判定する機能と、該判定結果に応じた文書認識処理を行う機能をコンピュータに実現させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体であることを特徴としている。

【0032】請求項25記載の発明では、文書画像中の各文字領域から行を切り出す機能と、行の高さに対する行内の各矩形高さの割合のヒストグラムまたは行内の矩形の最大高さに対する行内の各矩形高さの割合のヒストグラムを基に前記各文字領域が日本語領域であるか英語領域であるかを判定する機能と、何れの領域にも判定できない不明領域については、予め前記ヒストグラムを基に日本語の特性値と英語の特性値とを算出する機能と、前記不明領域が前記何れの特性値に近いかを算出することによって日本語領域であるか英語領域であるかを判定する機能をコンピュータに実現させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体であることを特徴としている。

【0033】請求項26記載の発明では、文書画像から

縮小画像を生成する機能と、該縮小画像から文字領域を生成する機能と、該文字領域から外接矩形を生成する機能と、所定の位置関係にある矩形同士を統合する機能と、統合後の矩形の縦横比のヒストグラムを基に前記文字領域が日本語領域であるか英語領域であるかを判定する機能をコンピュータに実現させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体であることを特徴としている。

【0034】

【発明の実施の形態】以下、本発明の一実施例を図面を用いて具体的に説明する。

〈実施例1〉図1は、本発明の実施例1の構成を示す。図において、101は、文書画像を入力する画像入力手段、102は、入力文書画像を縮小する画像縮小手段、103は、文書画像から連結成分を抽出する連結成分抽出手段、104は、抽出した連結成分を分類し、統合することによって文字領域を生成する領域生成手段、105は、文字領域単位またはページ単位で日本語と英語を判別する日英判別手段、106は、全体を制御する制御部、107は、入力された文書画像データや連結成分データ、領域データなど各種データを記憶するデータ記憶部、108は、データ通信路、109は、ネットワーク、回線などを介してホストなどに接続するデータ通信手段である。

【0035】図2は、本発明の実施例1の全体の処理フローチャートを示す。以下、図2を参照しながら、本発明の処理動作を説明する。まず、画像入力手段101は、文書を読み取ることによって文書画像を得る（ステップ201）。この画像入力手段は、例えばスキャナ、ファックスなどであり、またデータ通信手段109を介してネットワーク経由で別の機器から画像を得るようにしてもよい。

【0036】次に、画像縮小手段102は、入力された文書画像を縮小する（ステップ202）。この処理は、例えば入力文書画像を1/8程度にOR縮小する処理である。すなわち、8×8画素を1画素に縮小するもので、64画素中に1つでも黒画素があれば縮小画素は黒画素とする処理である。

【0037】連結成分抽出手段103は、縮小画像から黒画素連結成分を抽出する（ステップ203）。領域生成手段104は、抽出した連結成分を分類し、統合して文字領域を生成する（ステップ204）。この領域生成方法として、例えば特開平6-20092号公報に記載された公知の方法を用いればよい。このとき、各文字領域を構成する連結成分の情報はデータ記憶部107に格納、保持する。

【0038】続いて、生成した文字領域について、日英判別手段105は日本語か英語かの判定を行う（ステップ205）。

【0039】ステップ202において画像をOR縮小す

ることにより、近傍の黒画素どうしが融合する。ここで英文においては単語間にはスペースが存在し、単語内の文字間には非常に狭いという特徴がある。一方、日本語においては、句読点の前後以外では文字間隔は大きくは変わらない。

【0040】図3は、英文、日本語文の画像例と、その外接矩形を示す。英文画像301を縮小し、連結成分を抽出した結果を外接矩形で表現したものが外接矩形302である（なお、縮小処理しているので外接矩形302は、本来画像301より小さくなるべきだが、ここでは同じサイズで表現している）。英文画像では、単語毎に融合して連結成分が構成される。

【0041】日本語画像303と305の例について、同様に縮小して連結成分を抽出し、その外接矩形で表現すると、それぞれ外接矩形304、306のようになる。

【0042】英文の場合は、単語を構成する文字の数がある程度一定であるので、縦横比が2倍から6、7倍程度となる外接矩形が多くなる特徴がある。一方、日本語の場合は、外接矩形304に示すように英文では現れにくい長い矩形が生じたり、逆に外接矩形306のように細かい矩形が多く生じる特徴がある。

【0043】そこで、上記した連結成分矩形を「短」、「中」、「長」の3種類に分類し、これを各文字領域について集計する。図4は、実施例1の日英判定の処理フローチャートを示す。図4の処理は各文字領域毎に行われる。矩形の分類は、行方向が横の場合には例えば、幅/高さが2以下で「短」、幅/高さが2から6で「中」、それ以上で「長」とする（ステップ401）。そして、文字領域中におけるこの分類結果を集計し（ステップ402）、文字領域毎に日本語か英語かを判定する（ステップ403）。ここで、「短」矩形の数をSCNT、「中」矩形の数をNCNT、「長」矩形の数をLCNTとすると、日英の判定は図8（ステップ403の詳細フローチャート）に示すように行われる。

【0044】まず、 $LCNT / (NCNT + SCNT) > Th1$ が成り立つかどうか調べる（ステップ801）。Th1は予め定めたしきい値であり、例えば0.3程度とする。この条件式が成り立てば、長矩形が十分に多いということであり、当該文字領域は日本語領域であると判定する（ステップ804）。

【0045】次に、ステップ801でNoと判定されたとき、 $NCNT / (LCNT + SCNT) < Th2$ が成り立つかどうかを調べる（ステップ802）。Th2も予め定めたしきい値であり、例えば3とする。この条件式が成り立てば、中矩形が少ないということであり、当該文字領域は日本語領域であると判定する（ステップ804）。いずれの条件も満たさない場合は、英語領域と判定される（ステップ803）。

【0046】〈実施例2〉上記した実施例1では、文字

領域単位で日英の判定を行っている。この場合、文字領域によっては文字数が非常に少ない場合がある。そのような場合は、矩形の数が十分に得られないので矩形数の比率で日英判定を行うことが難しくなる可能性がある。実施例2は、矩形の数が十分でない場合を考慮した実施例である。

【0047】図5は、実施例2の処理フローチャートを示す。日英判別手段105は、集計された領域内の矩形の数が十分であるか否か（つまり所定の閾値 Th 以上あるか否か）を調べ（ステップ501）、十分でない場合には、前掲した特開平6-150061号公報に記載されているOCRを利用した日英判別を行う（ステップ503）。この場合は、文字の数が少ないのでOCR処理を施しても処理時間の増大は少なくてすむ。そして、矩形の数が十分である場合には実施例1で説明した矩形長による日英の識別を行う（ステップ502）。

【0048】〈実施例3〉次に、ページ単位で日英識別を行う実施例3について説明する。図6、7は、実施例3に係るステップ205の詳細フローチャートを示す。図6に示す方法は、「短」、「中」、「長」矩形の数の集計を文字領域毎でなくページ全体について行い（ステップ601、602）、その結果を使用してページ単位に日英の判定を行う（ステップ603）。この日英の判定方法は、図8の処理フローチャートに従って行う。このときのしきい値 $Th1$ 、 $Th2$ は文字領域単位の処理の場合と異なるしきい値としてもよい。

【0049】図7に示す方法は、各文字領域毎に日英の判別を行い（ステップ702）、その結果を基に当該ページの日英判定を行う（ステップ703）。具体的には、日本語領域と判定された領域の数を Jn 、英語領域と判定された領域の数を En として、 $Jn > En$ なら日本語ページ、 $En > Jn$ なら英語ページと判定する。 $Jn = En$ の場合はリジェクトし、あるいは日英の何れかに判定してもよい。

【0050】〈実施例4〉上記した実施例とは異なる特徴を利用した日英識別方法について説明する。図9は、実施例4の構成を示す。実施例1と異なる点は、行切り出し部902と、ブロック抽出部903と、ブロック内文字種判別部904を設けている点である。他の構成要素は実施例1のものと同様である。図10は、実施例4の処理フローチャートを示す。

【0051】まず、行切り出し部902は、文書画像の文字領域から行の切り出しを行う（ステップ1001、1002）。領域生成処理として、特開平6-20092号公報記載の技術を使用した場合には、領域を抽出した段階で行情報が得られているので、これを用いればよく、また電子通信学会論文「周辺密度分布、線密度、外接矩形特徴を利用した文書画像の領域分割」（秋山他、1986年8月、Vol. J69-D No. 8）に記載されている射影を用いる方法を用いてもよい。

【0052】次に、ブロック抽出部903は、単語相当のブロックを抽出する（ステップ1003）。このブロック抽出方法として、本出願人が先に特開平8-34781号で提案した方法を用いればよい。すなわち、ブロック抽出部111は、行データ内部の外接矩形を検出し、その外接矩形をブロックデータにまとめる。このブロックデータにまとめる方法は、次の通りである。文字矩形の間隔（まだ一つの矩形が一字とは確定されていない。従って、漢字の場合、偏とつくりに分離したものがそれぞれ一つの矩形となる場合も多い）のヒストグラムを求める。図18は、抽出された文字矩形と、矩形間の距離を示す。図19は、矩形間隔のヒストグラムを示す。

【0053】このヒストグラムにおいて、最も距離の短いピークは、漢字の偏とつくりの間隔や、プロポーショナル英字の同一単語内の文字間距離に現れる傾向がある。これらを統合しても異なる文字種がブロックに入ることには少ないので、それらを統合することでブロックデータを形成する。この処理を行うことによってプロポーショナルの単語や一字が分離する（つまり偏とつくりからなる）漢字が一つに統合されることになる。

【0054】また、最も距離の長いピークは、単語間の距離、句読点と次の文字との距離に現れることが多い。これらは（特に単語間の距離は）文字種が変わる場合の境目に用いられることが多く、同一ブロックになることを避けた。そこで、最も距離の長いピーク値以上の距離の文字矩形については、同一ブロックにしないように処理する。

【0055】さらに、対象矩形の両隣の矩形との距離（ A 、 B ）を測定し、その差（ $A - B$ ）が所定の閾値以上のとき、長い方の距離の矩形同志は統合せず、短い方の距離の矩形を統合するように処理する。図20は、矩形間の間隔の差が大きい位置で矩形の統合を行わない場合を説明する図である。図20では、差が所定の閾値以上大きい位置で矩形の統合を行わないので、3つのブロックが形成される。このような処理を行うことによって、プロポーショナルの英文などで、単語間の距離が絶対的に近くても、文字間距離とは差があるはずであるので、一つの単語だけをまとめて統合できる。また、プロポーショナルフォントであっても日本語の漢字部分は比較的等間隔に配置されるので、日本語文をまとめる場合にも都合がよい。

【0056】上記したブロック抽出方法を用いることによって、英文の場合、日本語文書と違って単語と単語の間は半角相当のスペースで区切られるために、他の文字種と混合してブロックデータとなることが避けられる。

【0057】続いて、ブロック内文字種判別部904は、ブロック毎の日英判別を行う（ステップ1004）。これも前掲した出願の方法を用いればよい。つまり、ブロック内文字種判別部904は、上記処理によ

てブロック化されたまとまりが、日本語であるか、英数字であるかという文字種の判定を行う。ブロック内は同一文字種として判断する。この文字種の判定は次のように行う。すなわち、ブロック内の矩形の幅に対して、該矩形の垂直方向の黒ランの数または白黒反転回数が所定の閾値以上のとき日本語文字と識別し、抽出されたブロック内の矩形の垂直方向座標値を基に英字を識別する。図21(a)、(b)は、日本語と英字の場合の垂直方向ランの数の具体例を示す。英数字ではノイズがない理想的な場合、最大で“g”の文字で4つのランができる(図21(b))。従って、5つ以上のランがカウントされる場合は日本語とする。図21(a)に示す文字「像」の場合、垂直方向のランの数は、文字の下の数字で示すように変化する。

【0058】日英判別手段905は、ブロック毎の判別結果を集計して当該領域の日英判別を行う(ステップ1005)。ここで、日本語と判定されたブロックの数をJ CNT、英語と判定されたブロックの数をE CNT、不定と判定されたブロックの数をN CNTとする。図11は、ステップ1005の詳細のフローチャートである。J CNT*Th3>E CNTのときは日本語と判定し(ステップ1101、1105)、そうではなく、E CNT>J CNTのときは英語と判定する(1102、1104)。それ以外の場合はリジェクトとする(ステップ1103)。しきい値Th3は、例えば2とする。

【0059】〈実施例5〉上記した実施例4では、文字領域単位で日英の判定を行っている。この場合、文字領域によっては文字数が非常に少ない場合がある。そのような場合は、矩形の数が十分に得られないのでブロックの判別結果数の比率で日英判定を行うことが難しくなる可能性がある。実施例5は、ブロックの数が十分でない場合の実施例である。

【0060】図12は、実施例5の処理フローチャートを示す。日英判別手段105は、集計された文字領域内のブロックの数が十分であるか否か(つまり所定の閾値Th以上あるか否か)を調べ(ステップ1201)、十分でない場合には、前掲した特開平6-150061号公報に記載されているOCRを利用した日英判別を行う(ステップ1203)。この場合は、文字の数が少ないのでOCR処理を施しても処理時間の増大は少なくすむ。そして、ブロックの数が十分である場合には実施例4で説明したブロック毎の判別結果による日英の識別を行う(ステップ1202)。

【0061】〈実施例6〉実施例6は、実施例4の文字領域毎の日英判別を、ページ単位の日英判別に変更したものである。実施例6の処理フローチャートは、図6、7を用いる。

【0062】図6の処理においては、J CNT、E CNT、N CNTの集計を文字領域毎でなくページ全体について行い、その結果を使用して、前述した図11の処理

方法によって日英の判定を行う。このときTh3は文字領域単位の場合とは異なってもよい。

【0063】図7の処理においては、まず、各文字領域毎に判別し、その結果から当該ページの日英判定を行う。具体的には、日本語領域と判定された領域の数をJ n、英語領域と判定された領域の数をE nとして、J n>E nなら日本語ページ、E n>J nなら英語ページと判定する。J n=E nの場合はリジェクトとしてもいいし、日英の何れかにしてもよい。

【0064】〈実施例7〉実施例7では、文字領域毎またはページ単位で日英判別を行う際に、図13に示すように矩形長を利用する日英判別処理(ステップ1301)と、ブロック毎の判別結果を利用する日英判別処理(ステップ1302)によって、それぞれ日英の判別を行う。そして、それぞれの判別結果から最終的に日英に判別を行う(ステップ1303)。

【0065】両者共に日本語または英語と判定された場合には、最終結果はそのまま日本語または英語と判定すればよい。何れかがリジェクトと判定された場合には、リジェクトでない方の判定結果を最終結果とする。

【0066】両者の判定結果が、一方が日本語で、他方が英語で、その結果が一致しない場合には、以下のいずれかの判定をする。

(1) リジェクトとする。

(2) 両者の確信度を算出し、値の大きな方の結果を採用する。

矩形長を利用する判別方法の確信度としては、例えば $LCNT/(NCNT+SCNT)>Th1$ で、 $Th1=0.3$ の場合には $LCNT/(NCNT+SCNT)*2.5$ の値(ただし上限を1とする)

$NCNT/(LCNT+SCNT)<Th2$ で、 $Th2=3$ の場合には $(LCNT+SCNT)/NCNT*2.5$ の値(ただし上限を1とする)

$NCNT/(LCNT+SCNT)>Th2$ で、 $Th2=3$ の場合には $NCNT/(LCNT+SCNT)*0.33$ の値(ただし上限を1とする)

とする。

【0067】ブロック毎の判別結果を利用する判別方法の確信度としては、例えば

$J CNT*Th3>E CNT$ で、 $Th3=2$ の場合には、 $J CNT/(E CNT*3)$ の値(ただし上限を1とする)

$E CNT>J CNT$ の場合には、 $E CNT/J CNT*0.7$ の値(ただし上限を1とする)

とする。

【0068】〈実施例8〉図14は、実施例8の構成を示す。また、図15は、実施例8の処理フローチャートを示す。この実施例では、入力された文書のページ全体について、日英判別部1412は、前述した実施例3、6の方法を用いて、そのページが日本語であるか英語で

あるかの日英識別処理を行い（ステップ1501、1502）、その判別結果に基づいて選択部1403は英文文書認識部1404または日本語文書認識部1405を選択し、選択された言語の文書認識処理を行い（ステップ1504、1505）、その認識結果をディスプレイなどの出力部に出力する（ステップ1506）。

【0069】なお、日本語と英語とではその属性が異なることから、領域分割処理やフォント識別処理なども切り替えた方がよい場合がある。そこで、本実施例の文書認識部は、文字認識処理だけではなく、上記した領域分割処理やフォント識別処理も含まれている。

【0070】〈実施例9〉図16は、実施例9の構成を示し、図17は、実施例9の処理フローチャートを示す。実施例8と異なる点は、日英識別を文字領域毎に行う点である。そのために、領域分割部1602は、入力文書を文字領域に分割する（ステップ1701、1702）。ここで、領域分割部では、日英両方に適応できる領域分割方法を使用する。分割処理された後、日英判別部1603は文字領域毎に、例えば前述した実施例1の方法を用いて日英識別処理を行い（ステップ1704）、その判別結果に基づいて選択部1604は英文文書認識部1605または日本語文書認識部1606を選択し、選択された言語の文書認識処理を行い（ステップ1705、1706）、その認識結果をディスプレイなどの出力部1607に出力する（ステップ1707）。なお、実施例9の文書認識部では、文書認識処理の他にフォント識別処理も行う。

【0071】〈実施例10〉前述した各実施例は、黒画素連結成分や矩形長を特徴量として日本語と英語を判定している。しかし、黒画素連結成分を用いる判定方法は処理時間がかかり、また矩形長を利用する方法はリジェクトの発生が高くなることもある。なお、外接矩形の上辺、下辺の行内での相対位置の頻度分布のピーク位置を基に和文か英文かを識別する方法もあるが（特公平7-21817号公報を参照）、傾きがある文書が入力された場合には、頻度分布が大きく変化し、識別精度が低下してしまうという問題点がある。

【0072】そこで、本実施例では、行高さに対する、行内の外接矩形の高さのヒストグラムを用いて日本語と英語を識別することにより、文書画像の領域毎に精度よくかつ高速に日本語と英語を識別するものである。そして、上記した日英識別方法でも判別不可能な領域に対しては、別の方法を用いて日英識別を行う。

【0073】図22は、実施例10の構成を示す。また、図23は、実施例10の全体の処理フローチャートである。まず、画像入力手段2201は、文書を読み取ることによって文書画像を得る（ステップ2301）。この画像入力手段は、例えばスキャナ、ファックスなどであり、またデータ通信手段2207を介してネットワーク経由で別の機器から画像を得るようにしてもよい。

【0074】次に、領域生成手段2202は、文字領域を生成する（ステップ2302）。この領域生成方法として、例えば特開平6-20092号公報に記載された方法を用いればよい。次に、行切り出し手段2203は、文字領域から文字認識のための行の切り出しを行なう。つまり、文字の外接矩形を求め、それらを統合して行を生成する（ステップ2303）。日英識別手段2204は、生成した文字領域について日英識別を行なう（ステップ2304）。

【0075】日英の識別は以下のようにして行う。図27は、日英識別（ステップ2304）の詳細のフローチャートである。図24は、切り出された行と行内の外接矩形の一例を示す。まず、行高さに対する、行内の外接矩形高さの割合の頻度分布を算出する（ステップ2701、2702）。行高さを $lineheight$ 、矩形高さを $height$ とする。割合を $height\ rate = height * 100 / lineheight$ とする。また、図25のような傾きのある文書の場合は、より精度良く日英識別するために、行高さの代わりにその行の矩形の高さの最大値を $lineheight$ として用いてもよい。つまり、傾きのある入力文書については、行内矩形の最大高さに対する、行内各外接矩形高さの割合のヒストグラムを基に日英識別する。

【0076】上記した割合 $height\ rate$ が例えば80以上の場合の矩形数を $lcnt$ とし、 $height\ rate$ が例えば70以上80未満の場合の矩形数を $ncnt$ とし、 $height\ rate$ が例えば40以上70未満の場合の矩形数を $scnt$ とする。文字領域内のすべての矩形に対し、 $lcnt$ 、 $ncnt$ 、 $scnt$ を求める。

【0077】図26は、日本語文書と英語文書について調べた矩形数の一例を示す。一般に、日本語は $lcnt$ が大きく、英語は $scnt$ が大きいという傾向がある。そこで、所定の閾値 thJ 、 thE を設定し、 $lcnt / scnt > thJ$ のとき日本語と判定し（ステップ2703）、 $lcnt / scnt < thE$ のとき英語と判定する（ステップ2704）。それ以外のときは不明領域とする（ステップ2705）。

【0078】上記した不明領域に対して、統計的手法を用いて日英識別することができる。図28は、不明領域に対する詳細な処理フローチャートである。例えば、あらかじめ日本語領域と英語領域の特徴値 $lcnt$ 、 $ncnt$ 、 $scnt$ を正規化し、その平均値と共分散行列の逆行列を日本語、英語についてそれぞれ求める。そして、平均値と共分散行列の逆行列を用いて、日本語、英語のそれぞれについてマハラノビス距離を求める（ステップ2801、2802）。

【0079】日本語のマハラノビス距離を Dj 、英語のマハラノビス距離を De とすると、所定の閾値を Me 、 Mj とすると、 $Dj / De > Me$ のとき英語と判定

し(ステップ2803)、 $Dj/De < Mj$ のとき日本語と判定する(ステップ2804)。何れの条件にも満足しない場合は不明領域と判定する(ステップ2805)。なお、上記したマハラノビス距離の代わりに、平均値とのユークリッド距離やシティブロック距離を用いてもよい。

【0080】さらに不明と判定された領域に対して、英文認識の確信度を用いて日英識別を行う。図29は、ステップ2805の詳細な処理フローチャートである。英文認識で確信度を算出する(ステップ2901)。次いで、算出された確信度について、例えば60%以上の確信度をもつ単語の個数をGood、60%未満で確信度0でない単語の個数をBad、確信度が0の単語の個数をZeroとする(ステップ2902)。

【0081】日英識別の判定値をValueとすると、 $Value = Good / (Good + Bad + Zero)$ とし(ステップ2903)、Valueが所定の閾値 th_{eocr} を超えれば(ステップ2904)、英語と判定し、それ以下ならば日本語と判定する。

【0082】なお、Zeroに重み付けしてもよい。Zeroを例えばBadの3個分とすると、 $Value$ は、 $Bad = Bad + Zero \times 3$ であるから $Value = Good / (Good + Bad)$ となり、Valueが閾値 th_{eocr} を超えれば英語、それ以下ならば日本語と判定することもできる。このように、日英識別判定のための文字数が少ない領域でも、英文認識による確信度で日英識別しているので、精度よく領域単位の日英識別が行われる。

【0083】〈実施例11〉本実施例は、入力文書画像を縮小した画像から外接矩形を生成し、生成された矩形同士で適当な統合を行い、統合後の矩形長の縦横比のヒストグラムを用いて日英識別をより精度良く行なう実施例である。

【0084】図30は、実施例11の構成を示す。また、図31は、実施例11の全体の処理フローチャートである。上記した実施例と同様にして画像入力手段3001によって入力された文書画像は、画像縮小手段3002によって縮小される(ステップ3101、3102)。この処理は、例えば文書画像を1/4程度にOR圧縮(4×4画素を1画素に縮小し、16画素中に1つでも黒画素があれば縮小画像は黒とする)する。

【0085】次に、領域生成手段3003は、文字領域を生成する(ステップ3103)。この領域生成方法として、例えば特開平6-20092号公報に記載された方法を用いればよい。続いて、矩形統合手段3004は、日英の特性が良く表れるように、矩形の統合を行なう(ステップ3104)。例えば、図32に示すように、矩形1、2のy座標(縦方向)の上下座標が近くかつ、隣同士の矩形1、2のx座標が非常に近い場合(例えば、矩形間の水平距離が英語のスペースに相当する距

離より小さい場合)、矩形を統合する。また、例えば、図33に示すように、左側の矩形1が右側の矩形2をy座標で包含する位置関係にありかつ、隣同士の矩形1、2のx座標が非常に近い場合(例えば、矩形間の水平距離が英語のスペースに相当する距離より小さい場合)、矩形を統合する。

【0086】そして、矩形縦横比(矩形長縦/矩形長横)を用いて、長矩形、中矩形、小矩形、極小矩形の4つの特徴量に分ける(図34)。一般に、日本語は長矩形の出現する割合が高く、また、英語は中矩形の出現する割合が高い。この特性の違いを利用して、日英識別手段3005は、識別判定式を作成し、日英識別を行なう(ステップ3105)。図35は、日英識別処理の詳細のフローチャートである。

【0087】

例えば、領域内での長矩形の領域数 $lcnt$

領域内での中矩形の領域数 $ncnt$

領域内での小矩形の領域数 $scnt$

領域内での極小矩形の領域数 $sscnt$ (ノイズの場合が多い)を算出し(ステップ3501)、領域内での長矩形の割合 $ratio1 = lcnt / (ncnt + scnt)$ を算出し(ステップ3502)、領域内での中矩形の割合 $ratio2 = ncnt / (lcnt + scnt)$ を算出する(ステップ3503)。なお、上記割合を算出するとき、 $sscnt$ はノイズとして無視した。

【0088】そして、 $ratio1$ をx座標、 $ratio2$ をy座標とし、誤識別を極力少なく、日英重なっている部分はリジェクトになるように、日本語領域、英語領域、リジェクト領域に分ける。例えば、 $ratio2 / ratio1 > th_E$ ならば英語領域と判定(ステップ3504)し、 $ratio2 / ratio1 < th_J$ ならば日本語領域と判定し(ステップ3505)、それ以外の領域は日英不明とする(ステップ3506)。ここで、 th_E 、 th_J は所定の閾値である。

【0089】日英不明と判定された領域に対して、実施例10と同様に、統計的手法を用いて日英識別する。例えば、あらかじめ日本語領域と英語領域の特徴値 $lcnt$ 、 $ncnt$ 、 $scnt$ を正規化し、その平均値と共分散行列の逆行列を日本語、英語でそれぞれ求める。平均値と共分散行列の逆行列を用いて日本語、英語のそれぞれのマハラノビス距離を求める。日本語のマハラノビス距離を Dj 、英語のマハラノビス距離を De とすると、所定の閾値を Me 、 Mj とすると、 $Dj/De > Me$ のとき英語、 $Dj/De < Mj$ のとき日本語と判定する。何れの条件も満たさない場合は不明と判定する。なお、マハラノビス距離の代わりに、平均値とのユークリッド距離やシティブロック距離を用いてもよい。

【0090】〈実施例12〉本発明は上記した実施例に限定されず、ソフトウェアによっても実現することができる。本発明をソフトウェアによって実現する場合に

は、図36に示すように、CPU、メモリ、表示装置、ハードディスク、キーボード、CD-ROMドライブ、スキャナなどからなるコンピュータシステムを用意し、CD-ROMなどのコンピュータ読み取り可能な記録媒体には、本発明の日本語英語判定機能、文書認識機能を実現するプログラムなどが記録されている。また、スキャナなどの画像入力手段から入力された文書画像などは一時的にハードディスクなどに格納される。そして、該プログラムが起動されると、一時保存された文書画像データが読み込まれて、日本語英語判定処理、文書認識処理を実行し、その結果をディスプレイなどに出力する。

【0091】

【発明の効果】以上、説明したように、本発明によれば、複数の判定方法を併用しているので、高精度に日本語と英語とを判別することができる。また、文書画像中の文字領域毎に精度よく日本語と英語の判別を行うことができ、文書画像のページ単位に、精度よく日本語と英語の判別を行うことができる。さらに、日本語または英語と判定された文書画像に対して、適切な文書認識処理を実行しているため、高精度な認識結果を得ることができる。

【図面の簡単な説明】

【図1】本発明の実施例1の構成を示す。

【図2】本発明の実施例1の全体の処理フローチャートを示す。

【図3】英文、日本語文の画像例と、その外接矩形を示す。

【図4】実施例1の日英判定の処理フローチャートを示す。

【図5】実施例2の処理フローチャートを示す。

【図6】実施例3に係るステップ205の第1の詳細フローチャートを示す。

【図7】実施例3に係るステップ205の第2の詳細フローチャートを示す。

【図8】ステップ403の詳細フローチャートを示す。

【図9】実施例4の構成を示す。

【図10】実施例4の処理フローチャートを示す。

【図11】ステップ1005の詳細のフローチャートである。

【図12】実施例5の処理フローチャートを示す。

【図13】実施例7の処理フローチャートを示す。

【図14】実施例8の構成を示す。

【図15】実施例8の処理フローチャートを示す。

【図16】実施例9の構成を示す。

【図17】実施例9の処理フローチャートを示す。

【図18】抽出された文字矩形と、矩形間の距離を示す。

【図19】矩形間隔のヒストグラムを示す。

【図20】矩形間の間隔の差が大きい位置で矩形の統合を行わない場合を説明する図である。

【図21】(a)、(b)は、日本語と英字の場合の垂直方向ランの数の具体例を示す。

【図22】実施例10の構成を示す。

【図23】実施例10の全体の処理フローチャートである。

【図24】切り出された行と行内の外接矩形の一例を示す。

【図25】文書が傾いている場合の行と行内の外接矩形の一例を示す。

【図26】日本語文書と英語文書について調べた矩形数の一例を示す。

【図27】日英識別(ステップ2304)の詳細な処理フローチャートである。

【図28】不明領域に対する詳細な処理フローチャートである。

【図29】ステップ2805の詳細な処理フローチャートである。

【図30】実施例11の構成を示す。

【図31】実施例11の全体の処理フローチャートである。

【図32】矩形を統合する例を示す。

【図33】矩形を統合する他の例を示す。

【図34】4種類に分類された矩形を示す。

【図35】実施例11の日英識別処理の詳細な処理フローチャートである。

【図36】実施例12の構成を示す。

【符号の説明】

- 101 画像入力手段
- 102 画像縮小手段
- 103 連結成分抽出手段
- 104 領域生成手段
- 105 日英判別手段
- 106 制御部
- 107 データ記憶部
- 108 データ通信路
- 109 データ通信手段

**This Page is Inserted by IFW Indexing and Scanning
Operations and is not part of the Official Record**

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ **BLACK BORDERS**
- ☐ **IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**
- ☐ **FADED TEXT OR DRAWING**
- ☒ **BLURRED OR ILLEGIBLE TEXT OR DRAWING**
- ☐ **SKEWED/SLANTED IMAGES**
- ☐ **COLOR OR BLACK AND WHITE PHOTOGRAPHS**
- ☐ **GRAY SCALE DOCUMENTS**
- ☐ **LINES OR MARKS ON ORIGINAL DOCUMENT**
- ☐ **REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**
- ☐ **OTHER:** _____

IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.